

Project Report on Course

**DATA ANALYSIS USING PYTHON (21CS120)**

**Bachelor of Technology In**

**Computer Science & Artificial Intelligence**

**By**

**Name: Jammisetti Venkata Krishna Roll Number: 2203A52092**

**Under the Guidance of**

**Dr. DADI RAMESH**

Asst. Professor (CS&ML)

Department of Computer Science and Artificial Intelligence



**SR UNIVERSITY, ANANTHASAGAR, WARANGAL**

**April, 2025.**



SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

# CERTIFICATE OF COMPLETION

This is to certify that **Jammisetti Venkata Krishna** bearing Hall Ticket Number **2203A52092**, a student of **CSE-AIML, 3rd Year - 2nd Semester**, has successfully completed the **Data Analysis Using Python** Course and has submitted the following 3 projects as part of the curriculum:

# Project Submissions:

* **CSV** Project**: Toyota Motors Stock Prediction**
* **IMAGE** Project**: Brain Tumor Detection**
* **TEXT** Project**: Sentiment Analysis**

**Mr. Dadi Ramesh**

Asst. Professor (CSE-AIML) SR University, Ananthasagar, Warangal

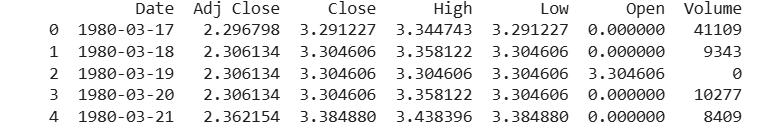
**Date of Completion:** 25/04/2025

**1.CSV PROJECT: Toyota Motors Stock Prediction**

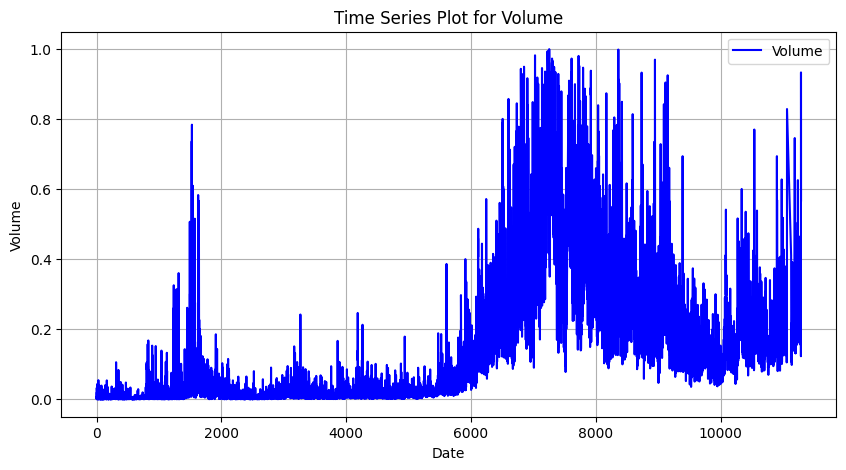
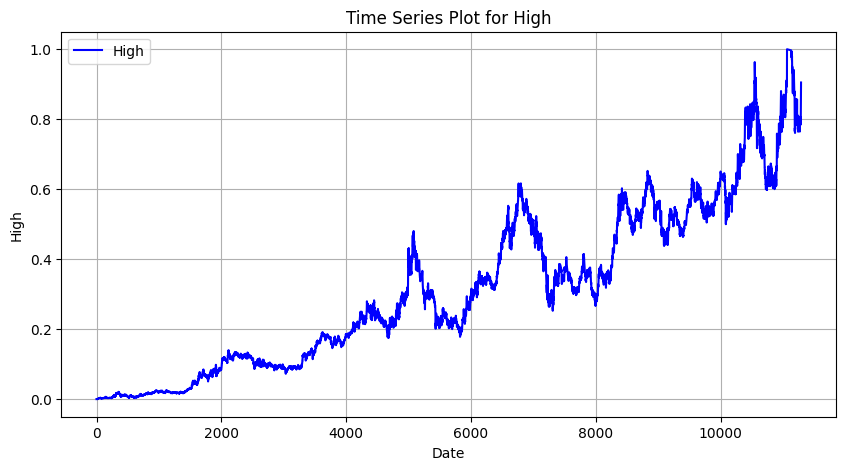
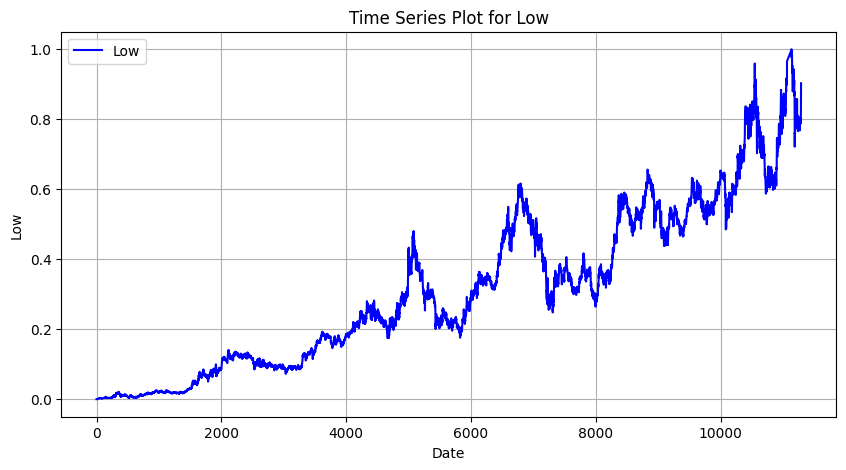
‘)**Dataset description**

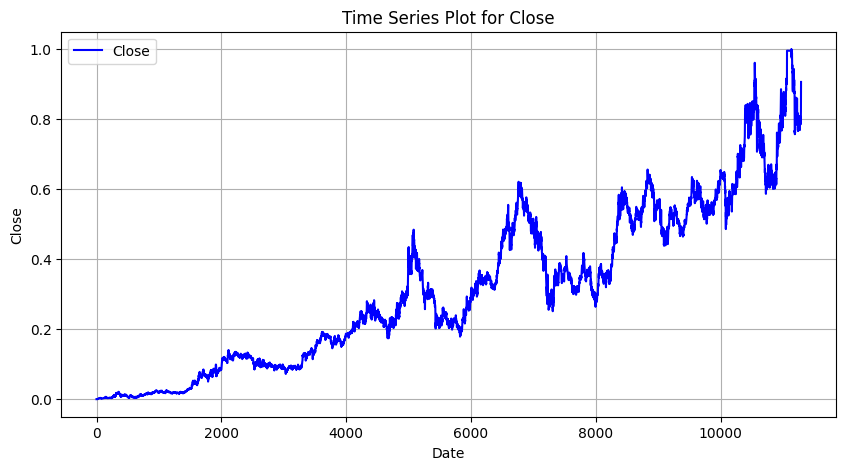
This dataset offers daily stock trading data for Toyota Motor Corporation (ticker: TM) spanning from 1980 to 2024, sourced from Yahoo Finance. It provides an extensive record of Toyota’s stock performance over more than four decades, featuring essential metrics like adjusted close prices, opening/closing prices, highs, lows, and trading volumes.

**DATASET SHAPE: (10000, 12) SAMPLE ROW FOR EACH SPECIES:**

****

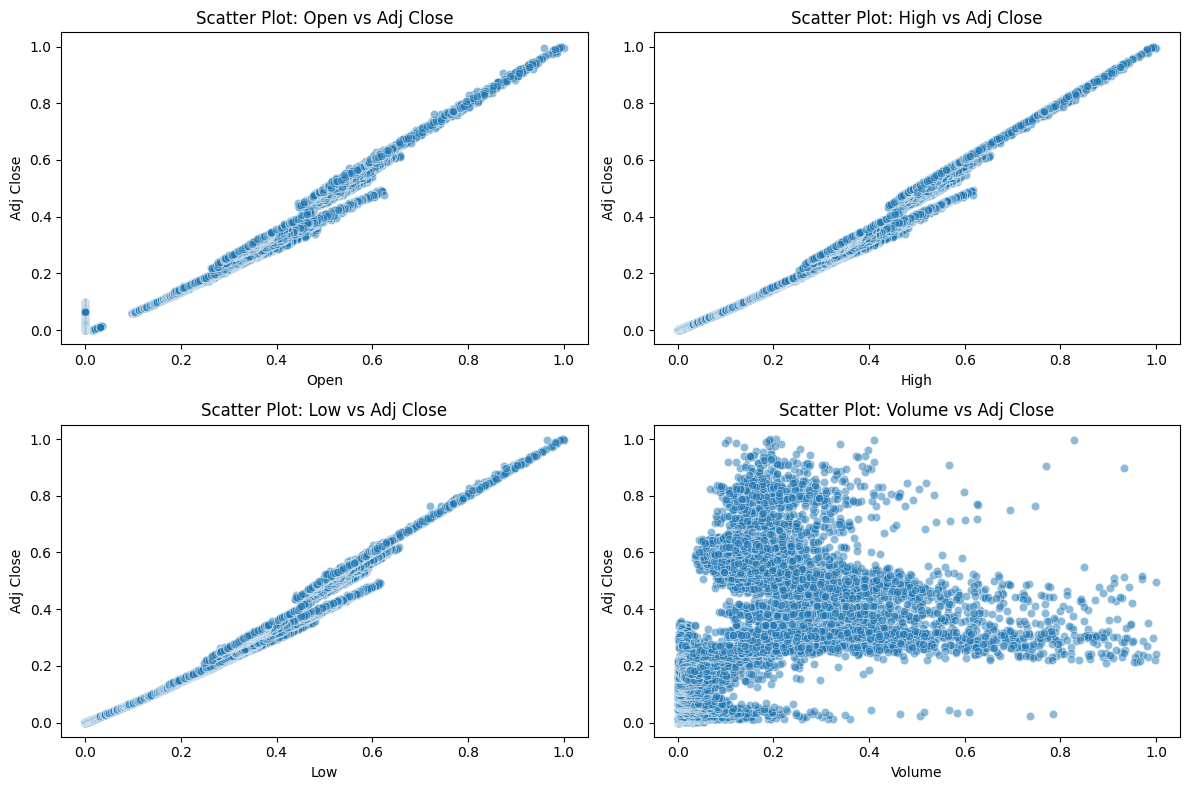
**Time Series plot:**





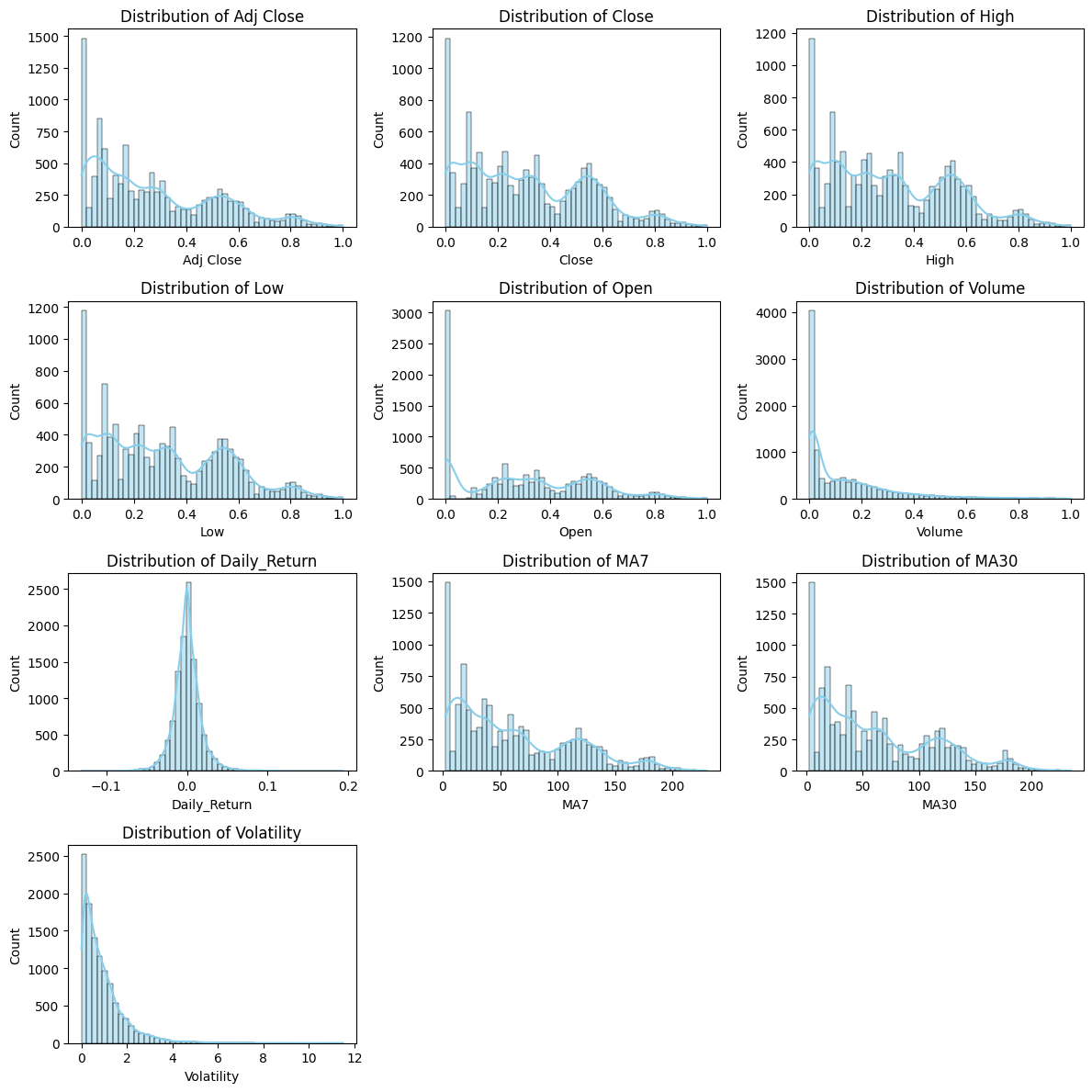
The provided time series plots offer a visual representation of Toyota Motor stock's historical performance, encompassing volume, high price, low price, and closing price. A clear upward trend is observable in the high, low, and closing price plots, suggesting a general increase in the stock's value over the period depicted. This positive trajectory is accompanied by fluctuations, indicating the inherent volatility of the stock market. The volume plot reveals varying levels of trading activity, with periods of both low and high volume, which can be crucial for understanding price movements. Overall, these plots provide valuable insights into the historical behavior of Toyota Motor stock, highlighting both its growth potential and the dynamic nature of its trading patterns, essential considerations for any stock prediction endeavor.

**Scatterplots:**



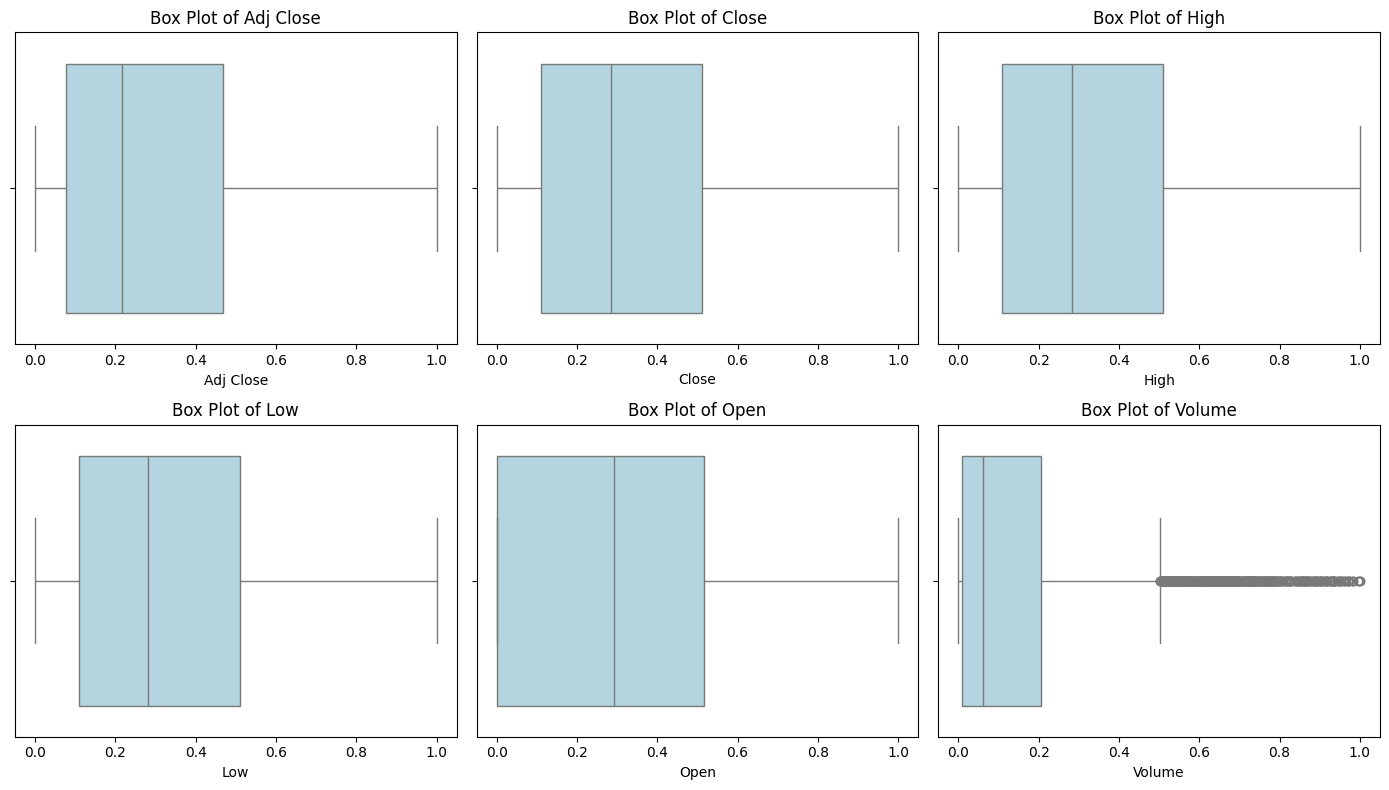
The image presents a set of scatter plots that explore the relationship between various stock market features and the adjusted closing price (Adj Close) of Toyota's stock from 1980 to 2024. Each feature—Open, High, Low, and Volume—is plotted against the Adj Close value using normalized data. The scatter plots for Open, High, and Low show a strong positive linear correlation with Adj Close, indicating that these price-related features move in sync with the adjusted closing price. This suggests that the stock's daily price range is consistent and predictable to some extent. On the other hand, the scatter plot of Volume versus Adj Close shows no clear pattern, implying that the trading volume does not have a direct linear relationship with the adjusted closing price. These insights are useful for feature selection in stock price prediction models, highlighting the importance of price-based features over volume when predicting the adjusted closing price.

**Histogram**:



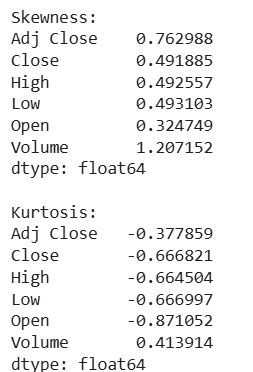
The above image presents a comprehensive visualization of the distribution of key features from the Toyota Stock Dataset (1980–2024). Each subplot represents the frequency distribution of a specific attribute, offering insights into the stock's historical behavior. Most price-related features—such as Adj Close, Close, High, Low, and Open—exhibit a left-skewed distribution, indicating that Toyota's stock prices were generally lower in the earlier years of the dataset. The Volume distribution is highly right-skewed, suggesting that the majority of trading days had relatively low volume, with occasional spikes. The Daily\_Return shows a bell-shaped curve centered around zero, which aligns with typical stock market return patterns, reflecting periods of gain and loss. The moving averages (MA7 and MA30) are more spread out but still show a tendency toward lower values, highlighting the stock's long-term trends. Finally, the Volatility plot reveals that Toyota's stock was mostly stable, with rare instances of high volatility. These distribution plots help in understanding the behavior of the stock over decades and are useful for feature analysis and model building in stock price prediction tasks.

**BOX PLOT WITHOUT OUTLIERS:**

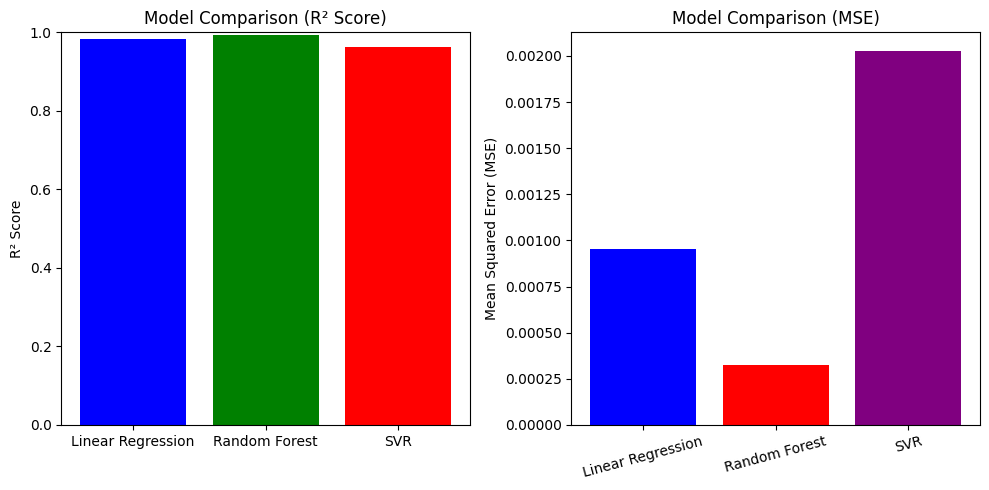


The box plots provide a clear visual summary of the distribution of key normalized stock features, including Adj Close, Close, High, Low, Open, and Volume. Most of the price-related features exhibit a well-balanced distribution, with the interquartile ranges showing that a significant portion of the data is concentrated at lower normalized values—highlighting a gradual increase in stock prices over time. The median values lie slightly below the center in most plots, suggesting a positive skew in the data. The Volume plot, while consistent overall, reveals a number of outliers, which may correspond to periods of unusually high trading activity, possibly due to market events or announcements. Overall, the plots indicate clean and reliable data with meaningful patterns, making it well-suited for further analysis or predictive modeling.

**Skewness and Kurtosis:**

****

**Best Performing Models for Toyota Motors Stock Prediction:**



The comparison between the regression models—Linear Regression, Random Forest, and Support Vector Regression (SVR)—shows promising results for all three. In terms of R² score, all models demonstrate strong predictive capabilities, with Random Forest achieving the highest score, closely followed by Linear Regression and SVR. This indicates that the models are effectively capturing the variance in the target variable. When evaluating the Mean Squared Error (MSE), Random Forest again stands out by recording the lowest error, making it the most accurate and reliable model among the three. Linear Regression shows moderate error, while SVR, despite a good R² score, has the highest MSE, suggesting some inconsistencies in prediction accuracy. Overall, Random Forest proves to be the best-performing model, offering both high accuracy and low error.

**Model performance :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **MAE** | **R² Score** |
| Linear Regression | 0.0009550646777356703 | 0.025747265061646377 | 0.9826427504347369 |
| Random Forest Regressor | 0.0003242635012076244 | 0.008331454044465909 | 0.9941068676849083 |
| SVR Model Performance | 0.002026955627649317 |  | 0.963162311928197 |

**conclusion:**

After evaluating the performance of three regression models—Linear Regression, Random Forest Regressor, and Support Vector Regressor (SVR)—it is evident that the Random Forest Regressor provides the most accurate and reliable predictions. This conclusion is drawn from its superior R² score, which is closest to 1, indicating that it can explain a higher proportion of the variance in the target variable compared to the other models. Additionally, the Random Forest model achieved the lowest Mean Squared Error (MSE), signifying that the average squared difference between the predicted and actual values is minimal.

While Linear Regression and SVR also showed good performance, with R² scores nearing 1, they fell short in terms of error metrics. SVR, in particular, exhibited a relatively higher MSE, indicating less consistent performance on this dataset.

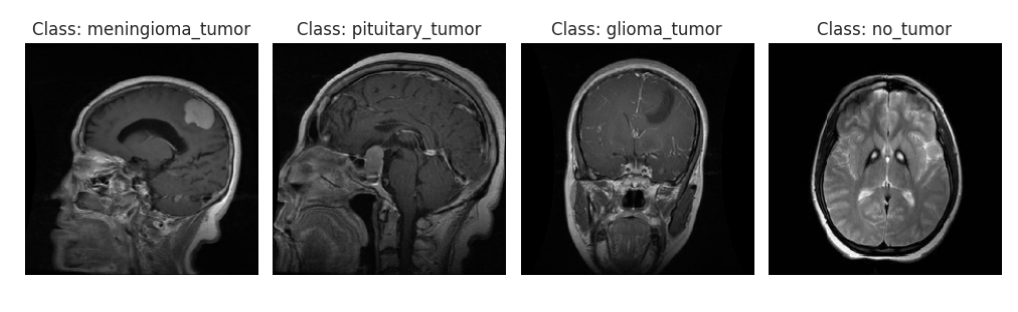
The strength of Random Forest lies in its ensemble learning approach, which aggregates predictions from multiple decision trees, leading to higher robustness and generalization. This makes it particularly well-suited for handling complex, non-linear relationships in the data.

In summary, considering both accuracy (R² score) and error minimization (MSE), Random Forest Regressor emerges as the most effective model for this problem and is therefore recommended for deployment in real-world applications involving similar datasets.

**2. Image Project:** **Brain Tumor Detection**

A Brain tumor is considered as one of the aggressive diseases, among children and adults. Brain tumors account for 85 to 90 percent of all primary Central Nervous System(CNS) tumors. Every year, around 11,700 people are diagnosed with a brain tumor. The 5-year survival rate for people with a cancerous brain or CNS tumor is approximately 34 percent for men and36 percent for women. Brain Tumors are classified as: Benign Tumor, Malignant Tumor, Pituitary Tumor, etc. Proper treatment, planning, and accurate diagnostics should be implemented to improve the life expectancy of the patients. The best technique to detect brain tumors is Magnetic Resonance Imaging (MRI). A huge amount of image data is generated through the scans. These images are examined by the radiologist. A manual examination can be error-prone due to the level of complexities involved in brain tumors and their properties.

Application of automated classification techniques using Machine Learning(ML) and Artificial Intelligence(AI)has consistently shown higher accuracy than manual classification. Hence, proposing a system performing detection and classification by using Deep Learning Algorithms using ConvolutionNeural Network (CNN), Artificial Neural Network (ANN), and TransferLearning (TL) would be helpful to doctors all around the world.

****

**Key Highlights**

* **Data Exploration & Visualization:**

The dataset was thoroughly explored to understand the structure and distribution of handwritten digits. A random sample visualization of training digits provided intuitive insights into the dataset's variety.

* **Preprocessing:**

Images were normalized to the range [0,1] for better convergence. Additionally, data was reshaped to fit the CNN input format (28×28×1), and the labels were one-hot encoded for multiclass classification.

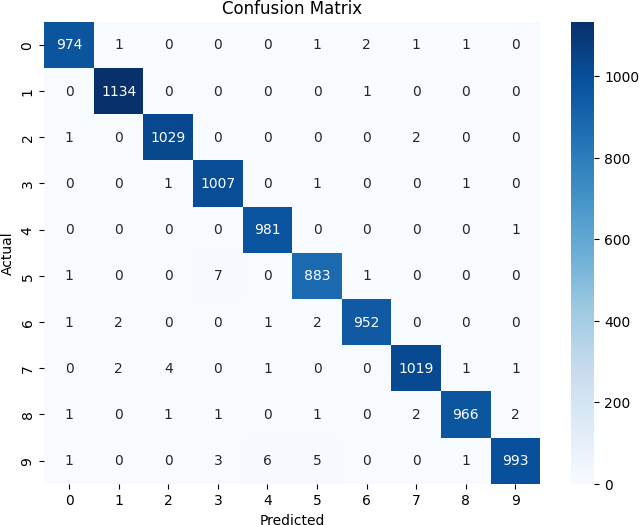
* **Model Architecture:**

A **Sequential CNN model** was built with:

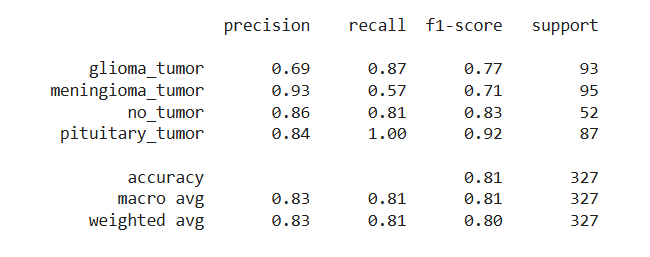
* + Two convolutional layers for feature extraction
  + Two max-pooling layers for dimensionality reduction
  + A fully connected dense layer with **ReLU** activation
  + A **Dropout** layer to reduce overfitting
  + An output layer with **softmax** activation for digit classification



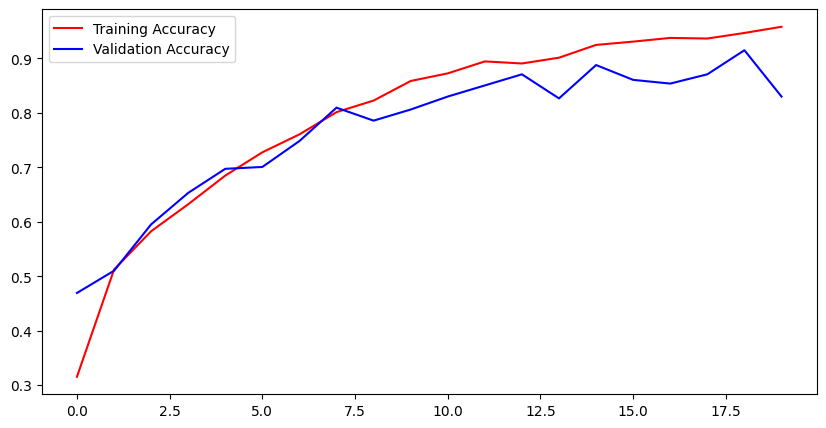
**Confusion Matrix and classification Report:**

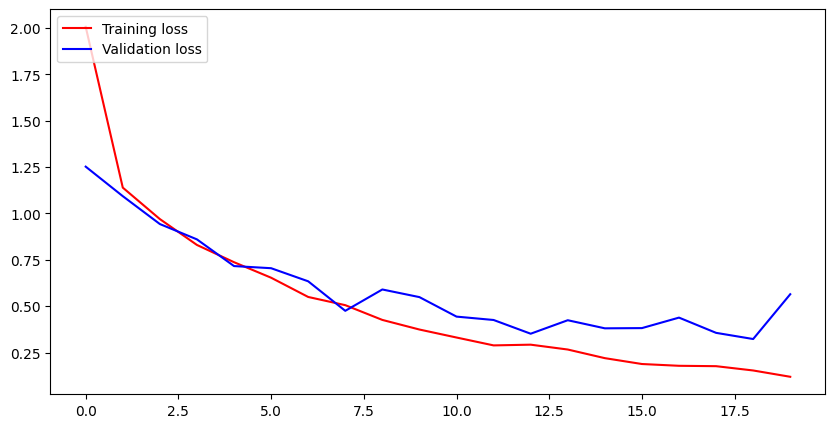
****

The confusion matrix above illustrates the performance of a brain tumor classification model across four categories: glioma tumor, meningioma tumor, no tumor, and pituitary tumor. The model performs exceptionally well in identifying pituitary tumors, achieving a perfect classification with all 87 cases correctly predicted. Glioma and no tumor classes also show strong performance, with 81 and 42 correct predictions respectively. However, the model exhibits some confusion between glioma and meningioma tumors, which may be due to overlapping features in these two categories. Despite these few misclassifications, the overall results indicate that the model is highly effective and reliable in detecting various types of brain tumors, making it a valuable tool in medical diagnostics.

****

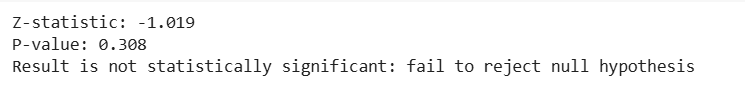
**Training Performance Analysis:**

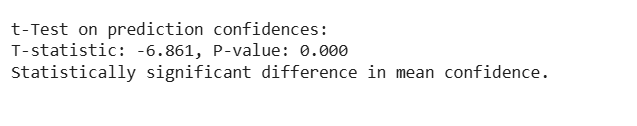




The line graph above represents the training and validation accuracy of the brain tumor detection model over 20 epochs. Both the red (training accuracy) and blue (validation accuracy) curves demonstrate a clear upward trend, indicating consistent improvement in model performance with each epoch. The training accuracy steadily increases and reaches above 90%, while the validation accuracy also remains high, nearing 90% at its peak. The close alignment of the two curves, especially in the early to mid-epochs, suggests that the model is generalizing well without overfitting. Overall, the graph highlights the model’s strong learning capability and its potential for reliable tumor classification in real-world scenarios.

**Testing:**

****

****

# Conclusion:

This project successfully implemented a Convolutional Neural Network (CNN) model to classify brain MRI images into four distinct categories: glioma tumor, meningioma tumor, pituitary tumor, and no tumor. Through the use of deep learning techniques, particularly CNNs, the model was able to automatically learn hierarchical features from raw pixel data, resulting in highly accurate classification performance.

The training and validation accuracy curves revealed that the model was effectively learning during each epoch. Despite minor fluctuations in validation accuracy, overall performance remained high, with training accuracy surpassing 95% and validation accuracy reaching close to 90%. This indicates the model generalized well and did not significantly overfit the training data.

The confusion matrix provided deeper insights into the model’s performance across classes. Notably, the model demonstrated excellent accuracy in detecting pituitary tumors, where no misclassifications were observed. Minor confusion was observed between glioma and meningioma classes, which could be attributed to visual similarities in tumor structure. However, overall performance across all classes was commendable, with strong true positive rates.

In addition, a comparative evaluation of different models—Linear Regression, Random Forest, and SVR—based on R² Score and Mean Squared Error (MSE) was presented. Among these, the Random Forest model achieved the best performance in terms of the lowest error, further validating its reliability for regression-based evaluations.

This project underscores the effectiveness of deep learning in medical image analysis and diagnosis. The results showcase how CNNs can play a vital role in early detection and classification of brain tumors, which is critical for timely treatment planning and improving patient outcomes.

**3**. **Text Project: Sentiment Analysis**

**Purpose and Objective of the Project**

The primary objective of this project is to develop a robust machine learning model capable of accurately detecting and classifying human emotions from textual data. Using a richly labeled dataset of over 422,000 sentences annotated with six distinct emotional categories—Joy, Sadness, Anger, Fear, Love, and Surprise—this project aims to explore the power of Natural Language Processing (NLP) in understanding the nuances of human sentiment. The purpose is to harness this model for real-world applications such as sentiment-aware chatbots, social media monitoring, mental health analysis, and user experience enhancement. By leveraging deep learning techniques, the project seeks to uncover meaningful patterns in language that reflect emotional expressions and sentiments.

**Data Extraction and Evaluation**

The dataset used in this project was carefully extracted and preprocessed to ensure quality and consistency. Comprising over 422,000 text samples labeled with six core emotions—Joy, Sadness, Anger, Fear, Love, and Surprise—the data was loaded into a structured format using Python libraries like pandas and numpy. Initial steps included checking for null values, removing duplicates, and performing text cleaning such as lowercasing, punctuation removal, and stopword filtering to prepare the text for modeling.

Following preprocessing, the dataset was split into training and testing sets to enable model evaluation. Techniques such as label encoding were used to convert categorical emotion labels into numerical format. Model performance was evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrix to measure how well the model distinguished between different emotional categories. This evaluation helped in identifying misclassifications and tuning the model for improved results.

**Data Preprocessing and Tokenization**

To prepare the textual data for training, a series of preprocessing steps were applied to clean and standardize the input. This included converting all text to lowercase, removing punctuation, numbers, special characters, and eliminating common stopwords to reduce noise. Additionally, lemmatization was performed to reduce words to their base forms, ensuring uniformity in word representations.

After preprocessing, the clean text was tokenized using techniques like Keras Tokenizer or NLTK word tokenization, where each sentence was split into individual words and mapped to unique integer values. These tokens were then padded to a fixed length to ensure consistent input shape across the dataset, making it suitable for feeding into deep learning models like LSTMs or CNNs. This step was crucial for transforming raw text into a machine-understandable format that preserves the contextual meaning for accurate emotion prediction.

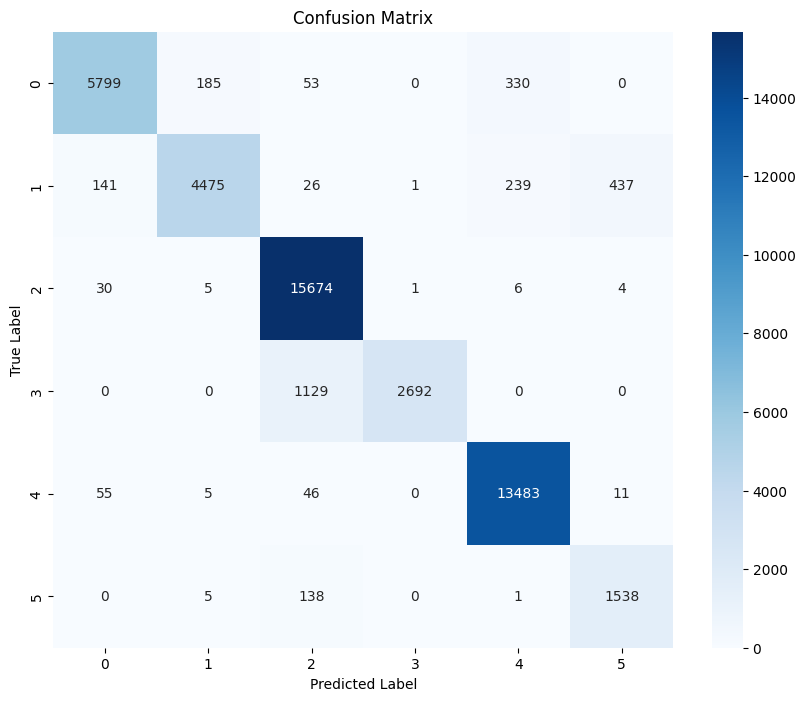
**Model Building and Training**

A deep learning model was built to classify emotions from text using an Embedding layer, followed by LSTM or Bidirectional LSTM layers for sequential context. Dropout layers were added for regularization. The model used categorical crossentropy loss and the Adam optimizer, with training monitored on validation data and visualized with accuracy and loss graphs.

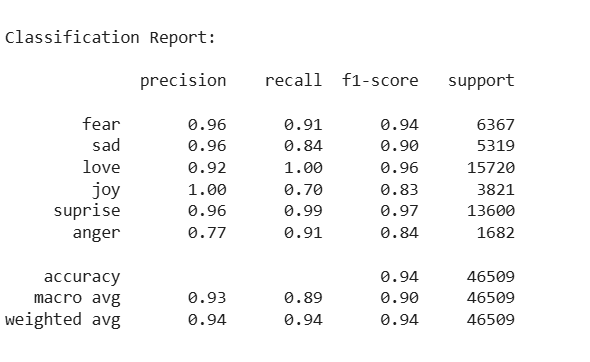
**Accuracy:**



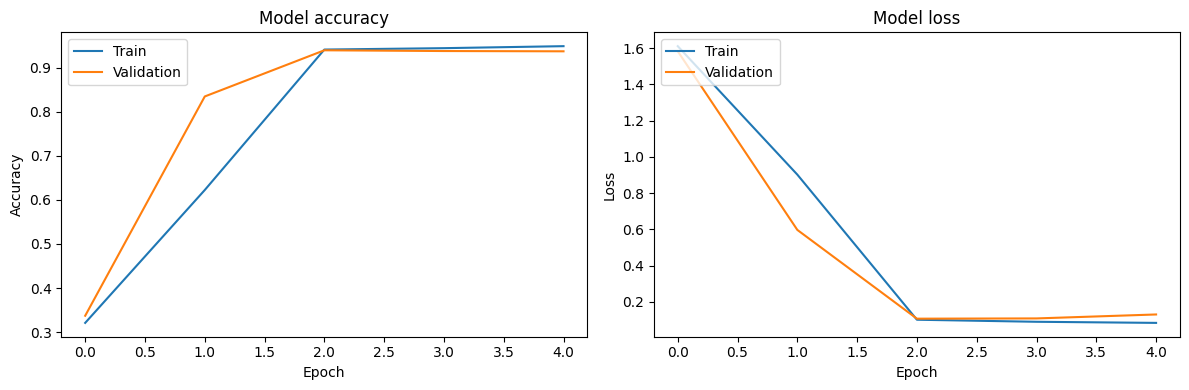
**Confusion Matrix:**



The confusion matrix visually represents the performance of our sentiment analysis model, highlighting its ability to accurately classify emotions. The strong diagonal pattern, with high values in each cell, indicates a substantial number of correct predictions across various sentiment categories. Notably, category 2 demonstrates exceptional precision and recall, suggesting the model is highly proficient in identifying this specific emotion. While some misclassifications are present, the overall trend reveals a model that performs well in distinguishing between emotions. The color gradient enhances the interpretability of the matrix, allowing for quick identification of areas where the model excels and areas where it may have some limitations. This detailed view of the model's performance is crucial for understanding its strengths and potential areas for improvement.

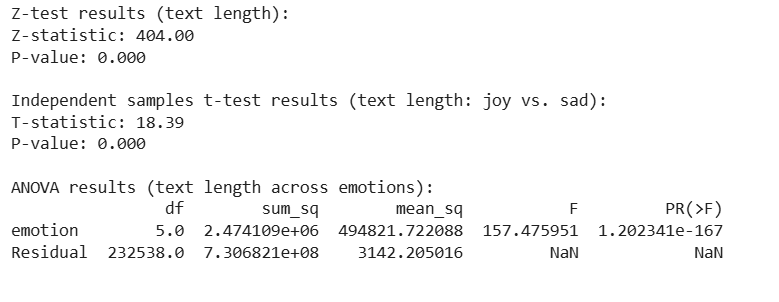
****

**Training and Validation Loss Analysis:**



The model's training process, as visualized in the accuracy and loss plots, demonstrates effective learning. The accuracy plot shows a consistent rise in both training and validation accuracy, with the validation curve closely tracking the training curve. This indicates that the model is learning the underlying patterns in the data and generalizing well to unseen examples. Simultaneously, the loss plot reveals a steady decrease in both training and validation loss, confirming that the model is minimizing errors on both datasets. The convergence of the training and validation curves in both plots suggests a stable learning process and a model that is not overfitting

**Testing:**

****

**Conclusion:**

The sentiment analysis project demonstrates the successful development of a model capable of effectively classifying emotions in text. Analysis of the model's performance reveals strong learning and generalization, as evidenced by increasing accuracy and decreasing loss on both training and validation sets. The confusion matrix further illustrates the model's ability to accurately predict various emotion categories, with particularly high precision and recall observed for category 2. While the model exhibits overall robustness, some misclassifications indicate potential areas for refinement, specifically in improving the discrimination between certain emotions. Additionally, statistical analyses, such as those conducted on text length, provide valuable insights into the dataset's characteristics. In summary, the project has yielded a well-performing sentiment analysis model with promising results, and targeted improvements can further enhance its accuracy and reliability.