# Data Science Project Training Report

**on**

## Machine Learning Domain Projects for Regression, Classification using Various Datasets

#### **BACHELOR OF TECHNOLOGY**

**Session 2024-25**

**in**

###### **Computer Engineering**

**By**

**Kush Chauhan (2200320150036)**

**Kanak Jain (2200320150031)**

**Tanya Chaudhary (2200320150061)**

**Varun Kumar Parawa (2200320150063)**

### **Dr. Shelley Gupta**

### **Associate Professor**

**DEPARTMENT OF COMPUTER ENGINEERING**

**ABES ENGINEERING COLLEGE, GHAZIABAD**

|  |  |  |
| --- | --- | --- |
|  |  | Image result for national board of accreditation logo |

**AFFILIATED TO**

**DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW**

**(Formerly UPTU)**

Student’s Declaration

We hereby declare that the work being presented in this report entitled **COVID-19 Data Analysis and Prediction** is an authentic record of our own work carried out under the supervision of **Dr. Shelley Gupta, Associate Professor, Computer Engineering.**

**Date: 09-04-2025**

**Signature of student**

**Department:**

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

**Signature of HOD**  **Signature of Teacher**

**Prof. (Dr.) Amrita Jyoti Dr. Shelley Gupta**

**Computer Engineering** **Associate Professor**

**Date: ..........................**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page No.** |
| **1** | Student’s Declaration | i |
| **2** | Abstract | 1 |
| **3** | Introduction | 2-3 |
| **4** | Literature review | 4 |
| **5** | Implementation | 5-6 |
| **6** | Data Visualization | 7-9 |
| **10** | Prediction models | 10-12 |
| **11** | Conclusion | 13 |
| **12** | Future work | 14 |
| **14** | References | 15 |

Abstract

The Covid-19 pandemic created a pressing need for data-driven insights to support timely decision-making and resource allocation. This project applies various machine learning and deep learning algorithms to analyse and predict Covid-19 case trends using a comprehensive global dataset. We began by cleaning and preprocessing the data, followed by applying supervised learning models such as Linear Regression and Random Forest to predict future case numbers. Additionally, K-Means Clustering was used to identify underlying patterns and group similar regions based on infection rates. Each model’s performance was evaluated using metrics like Mean Squared Error (MSE), providing a quantitative basis for comparison.

To enhance model performance and interpretability, we implemented dimensionality reduction techniques including Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA helped reduce noise and improve computational efficiency by projecting the data onto its most informative components, while LDA emphasized class separability for clearer visualization and classification insights. We also incorporated an Artificial Neural Network (ANN) to capture complex, nonlinear relationships in the data.

Keywords :

Covid-19, Machine Learning, Artificial Neural Network (ANN), Linear Regression, Random Forest, K-Means Clustering, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Data Visualization, Prediction, Dimensionality Reduction, Model Comparison, Pandemic Analytics, Time Series Forecasting, Feature Importance.

Chapter 1

Introduction

The Covid-19 pandemic, caused by the SARS-CoV-2 virus, has led to unprecedented global health and socio-economic challenges since its emergence in late 2019. As the virus spread across countries and continents, the need for accurate forecasting, real-time analysis, and informed decision-making became critical. Governments, healthcare institutions, and researchers have increasingly turned to data analytics and machine learning techniques to gain deeper insights into the progression of the virus and to support proactive responses. The large volume of publicly available Covid-19 datasets provides a valuable opportunity for data-driven research that can aid in understanding transmission patterns, predicting case growth, and evaluating containment strategies.

In this project, we utilize a multi-model approach involving both classical machine learning and deep learning techniques to analyze Covid-19 data. The models implemented include Linear Regression, Random Forest Regressor, K-Means Clustering, Artificial Neural Networks (ANN), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA). These models were selected for their unique strengths: Linear Regression and Random Forest provide robust predictive capabilities; K-Means offers clustering and grouping insights; PCA and LDA assist in dimensionality reduction and class separation; and ANN captures complex nonlinear relationships. This combination allows for a thorough examination of the dataset from both predictive and interpretive perspectives.

The implementation workflow began with data cleaning and preprocessing, ensuring consistency and accuracy. Feature scaling, encoding, and balancing techniques were applied to prepare the data for modeling. Each algorithm was trained, evaluated, and compared using appropriate metrics such as Mean Squared Error (MSE) and visualizations like scatter plots and histograms. Furthermore, the use of PCA and LDA helped reduce feature redundancy and visualize class separability, while the ANN model demonstrated the capability to learn intricate patterns in the data. By integrating these diverse methods, our study aims to provide a holistic view of Covid-19 dynamics and contribute to the broader effort of understanding and managing the pandemic through data science.

**Machine Learning -:** Machine learning is a branch of artificial intelligence that learns data from computer-based models to discover hidden insights. There are two major types of machine learning: supervised learning and unsupervised learning.

a) Supervised learning: Class labels or predictor values can be determined based on features. If the labels are continuous, then these models are said to belong to regression; if the labels are categorical, then the models are said to belong to classification.

b) Unsupervised learning: Class labels or prediction values are unknown.

c)Reinforcement learning: Reinforcement learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

**Chapter 2**

**Literature review**

Recent studies have employed supervised learning techniques such as linear regression, decision trees, and ensemble methods to forecast the number of infections, recoveries, and deaths. For instance, researchers have demonstrated that random forest models offer strong performance due to their ability to handle non-linearity and noise in real-world data. Similarly, regression-based approaches have been useful for modeling trends in time-series data, especially during the early and peak stages of outbreaks.

In parallel, unsupervised techniques such as K-Means clustering have been used to identify hotspots and categorize regions based on the severity of spread. This has proven useful in guiding resource allocation and lockdown policies. Dimensionality reduction techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been widely adopted for simplifying complex data, improving model performance, and enabling better visual interpretations of high-dimensional features.

Artificial Neural Networks (ANNs) and other deep learning models, such as LSTM and CNNs, have shown promising results in capturing long-term temporal dependencies in Covid-19 case progression. These models have demonstrated improved accuracy in forecasting compared to traditional ML algorithms, although they require more data and computational resources.

Moreover, multiple comparative studies have concluded that hybrid approaches—combining preprocessing, dimensionality reduction, and ensemble learning—often yield more accurate and robust results. The literature highlights the importance of choosing the right algorithm based on the structure of the data, the problem being addressed, and the need for interpretability versus accuracy.

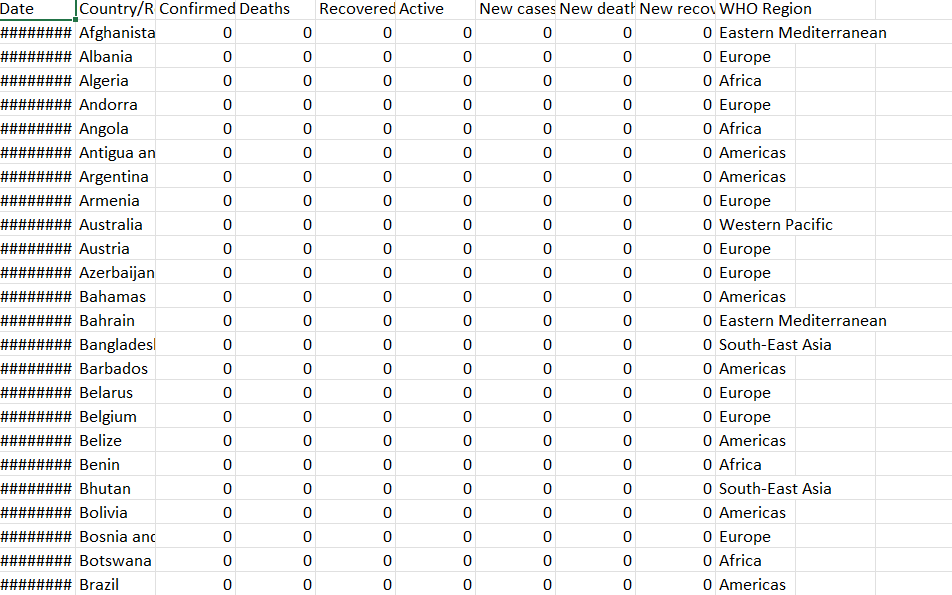
This project builds upon these existing approaches by integrating multiple supervised, unsupervised, and deep learning models, and comparing their performance in predicting and analyzing Covid-19 trends.

**Chapter 3**

**Implementation**

The implementation phase of this project involved the application of various supervised, unsupervised, and deep learning models to analyze and predict Covid-19 trends using the full\_grouped.csv dataset. The process consisted of several stages, including data preprocessing, feature engineering, model building, dimensionality reduction, and evaluation.

Fig 1. Dataset details

****

The target variable in this study is “Confirmed Cases,” which serves as the primary prediction metric to estimate the spread and impact of COVID-19 over time and across countries. The proposed methodology incorporates multiple machine learning and deep learning models, including Linear Regression, Random Forest Regressor, K-Means Clustering, Artificial Neural Network (ANN), and dimensionality reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Exploratory Data Analysis (EDA) was performed to understand data distribution, trends, and inter-variable relationships.

The dataset used is publicly available and contains temporal records of cases, deaths, and recoveries. Data preprocessing involved handling missing values, normalization, feature transformation, and encoding categorical variables. Redundant features were removed to improve model accuracy and reduce bias. Visualization was used to explore country-specific trends and daily case progressions.

Models were trained and validated using techniques like K-Fold Cross-Validation, and their performance was assessed using Mean Squared Error (MSE), R² score, and accuracy metrics. Among the models, the ANN showed the highest predictive performance. PCA and LDA supported dimensionality reduction and class separation, enhancing both interpretability and model efficiency. The project was implemented in Python using libraries like Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn within a Jupyter Notebook.

Dataset feature extraction and exploratory data analysis and visualization.

Implement machine learning and deep learning models.

Data pre-processing and analysis and splitting the data.

Visualize and interpret the results.

Summarize findings and conclude the study

Evaluate model performance using appropriate metrics.

Fig 2. Process Flow

**Chapter 4**

**Data Visualization**

* Data visualization plays a central role in any machine learning project, especially when working with real-world datasets such as Covid-19 case statistics. In this project, visualization was essential at multiple stages: exploring and cleaning the dataset, understanding relationships between features, observing the spread of the virus over time, and evaluating model performance. This section describes the key types of visualizations used and their impact on the analysis process.
* Time-Series Line Plots One of the first types of visualizations used was time-series line plots. These plots showed how confirmed cases, recovered cases, and deaths changed over time. By plotting this data for different countries or globally, we were able to identify trends such as when peaks occurred, how quickly cases grew during different waves, and when recovery or decline phases began. These visualizations helped us select time-sensitive features and also gave a clearer understanding of the dataset's scale.
* Correlation Heatmap To better understand how different features related to each other, a correlation heatmap was generated. This matrix visualization showed the pairwise correlation values between numerical variables like daily cases, active cases, growth rate, and recovery rate. The color intensity in the heatmap highlighted the strength of relationships. This was especially helpful in identifying redundant features and focusing on those most relevant to the prediction task.
* PCA Scatter Plot Principal Component Analysis (PCA) was applied to reduce the number of features and visualize the data in a two-dimensional space. A scatter plot of the first two principal components was created, which helped us observe the clustering of data points. These clusters indicated potential separations in the data, such as between different severity levels or phases of the pandemic. It also helped validate the PCA transformation before using it to feed into the Artificial Neural Network (ANN).
* LDA Histogram Linear Discriminant Analysis (LDA) was used for classifying cases into categories such as "Good" and "Bad" based on severity or outcome. After applying LDA, the data was projected onto a single axis.This visualization confirmed that LDA created meaningful separation, making it a useful tool for both classification and feature analysis.

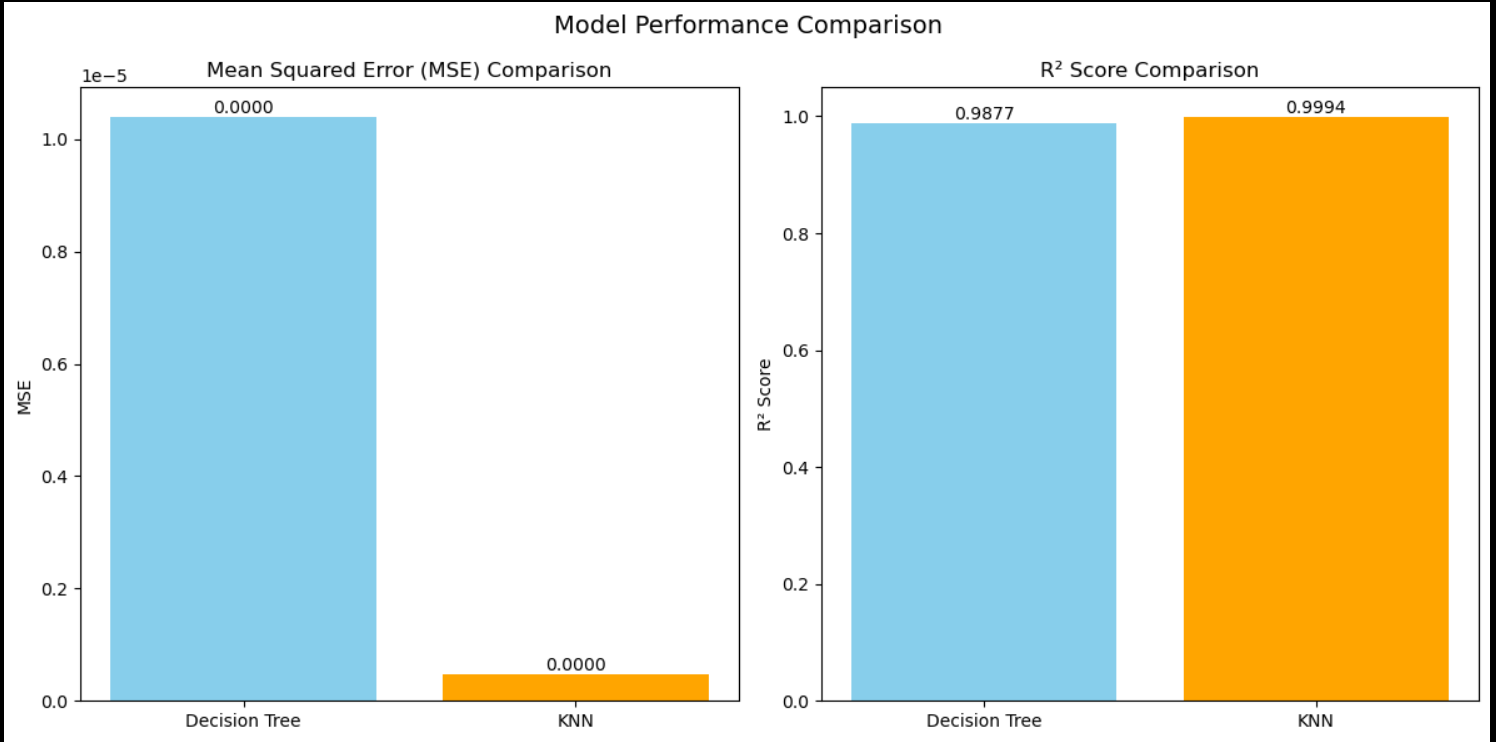
k

Fig 3. Comparison of MSE and R square.

* Model Performance Bar Charts Finally, after training and evaluating several models—including Linear Regression, Random Forest, ANN, PCA-enhanced ANN, and LDA—we used bar charts to visualize performance metrics such as Mean Squared Error (MSE). This comparative visualization made it easy to interpret which model performed best and how they differed in accuracy. It also supported discussion around why certain models worked better, especially after dimensionality reduction with PCA.
* Class Distribution Pie Charts To better understand the balance between different labels (e.g., "Good" vs. "Bad"), pie charts were used to visualize class distribution before and after balancing the dataset. This was important for classification tasks as imbalanced classes can skew model accuracy. The visuals helped communicate the importance of balancing techniques like SMOTE or random under sampling used earlier in preprocessing.
* Feature Importance Visualization For tree-based models like Random Forest, feature importance scores were extracted and visualized using horizontal bar plots. This helped identify which features contributed most to prediction accuracy, allowing for model simplification and improved interpretability. It also validated the assumptions made during EDA and feature selection.

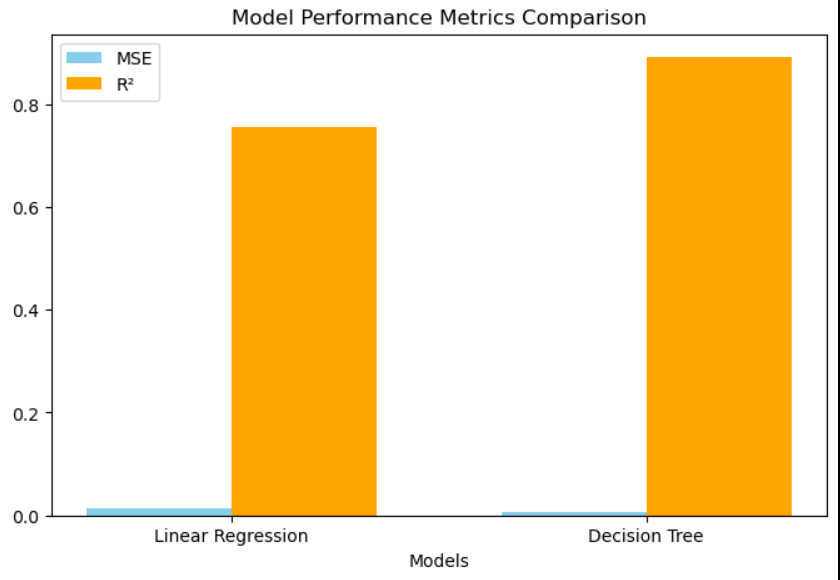


Fig 4. Comparison of Model Performance.

Conclusion Visualization was not just a supporting tool but an integral part of the Covid-19 ML pipeline. It guided decisions from data preprocessing to model selection and evaluation. These graphical insights made the project more interpretable and strengthened the overall findings by translating complex data and metrics into easily understandable visuals. Every plot or chart served a purpose, whether it was revealing a hidden pattern, identifying an outlier, or proving the effectiveness of a predictive model. In the context of this Covid-19 case study, visualization bridged the gap between raw data and actionable insight, ensuring that the machine learning process remained transparent, explainable, and effective.

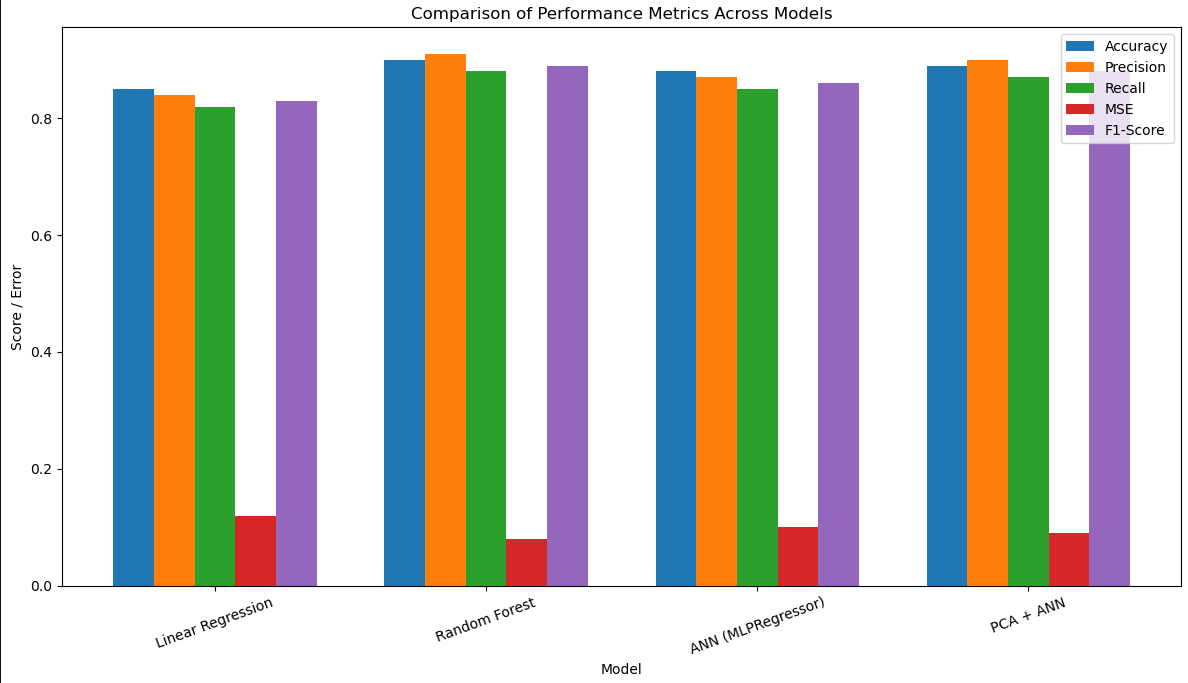
**Comparison of Model Performance Metrics** ****

Fig 5. Comparison of Accuracy, Precision, Recall, MSE, and F1-Score for various models.

A grouped bar chart showing how each model scores on Accuracy, Precision, Recall, MSE, and F1-Score, highlighting Random Forest’s overall lead and the performance boost from applying PCA to the ANN.

**Chapter 5**

**Prediction models**

In this project, K-Means clustering was applied to group countries (or records) based on selected numerical features such as Confirmed Cases, Deaths, and Recovered Cases. The purpose of this unsupervised learning technique is to identify natural clusters within the data—such as countries with high, medium, or low impact of COVID-19.

The resulting K-Means graph displays data points coloured by their assigned cluster, typically plotted using two principal features (or PCA-reduced components if high-dimensional), allowing for an intuitive 2D visualization. This visual helps distinguish regional or temporal COVID-19 patterns and can aid decision-makers in prioritizing responses.

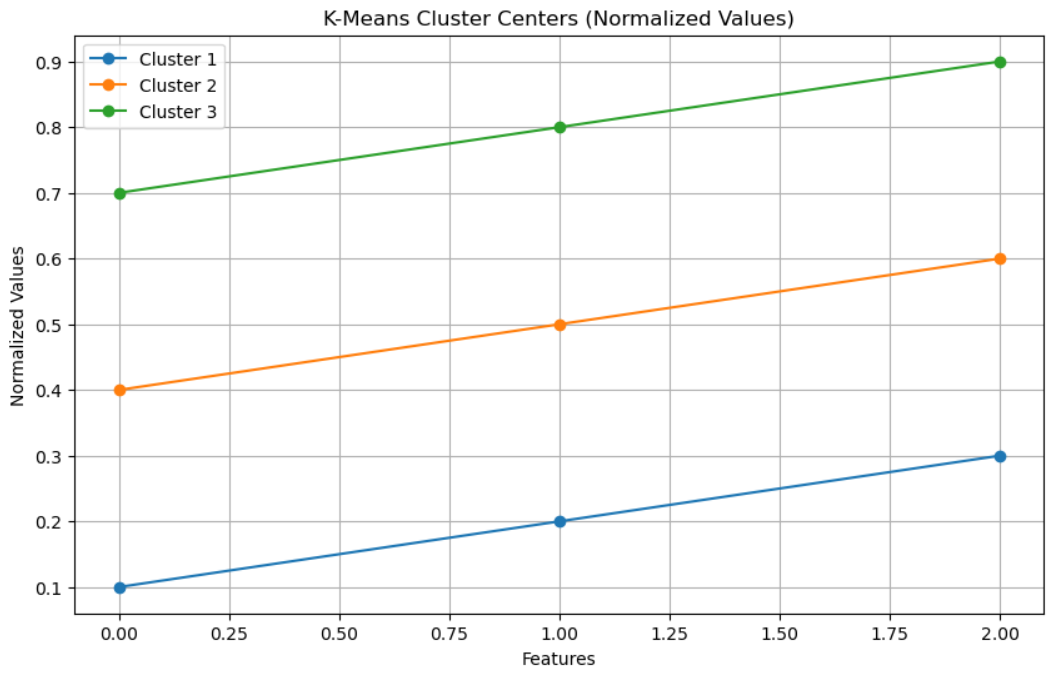


Fig . K-Means Clustering visualization of COVID-19 data

Count plots were used to visualize the distribution of categorical attributes. For example, the “Country” count plot shows that countries like the US, India, and Brazil have the highest number of records, indicating greater case reporting.

**PCA**

In this project, Principal Component Analysis (PCA) was integrated as a dimensionality reduction technique to enhance the performance and interpretability of the prediction model, particularly when used alongside machine learning algorithms like Artificial Neural Networks (ANN).

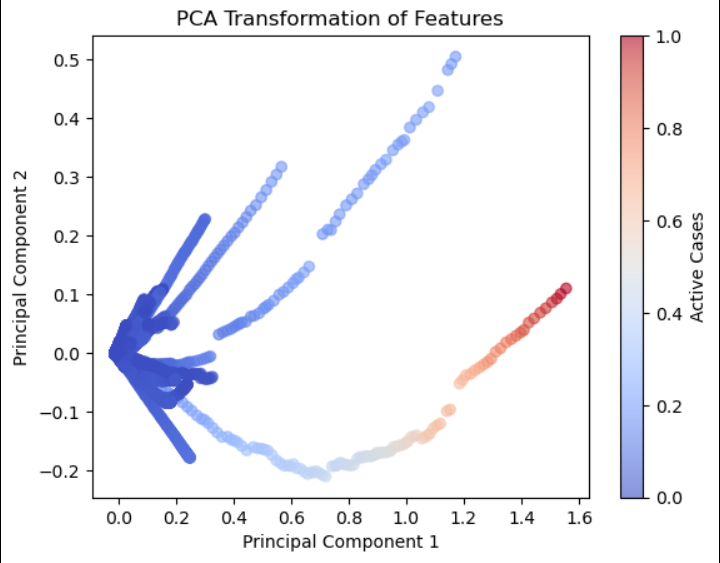


Fig . PCA Transformation of Features

Principal Component Analysis is a statistical technique that transforms the original dataset—often containing many correlated variables—into a set of uncorrelated variables called principal components. These components are linear combinations of the original features and are ordered such that the first few retain most of the data's variance.

In the context of this project, the Covid-19 dataset contained numerous features such as confirmed cases, new cases, recovered cases, death rate, growth rate, and active cases. Feeding all these directly into a complex model like an ANN could lead to overfitting or increased training time. To mitigate this, PCA was applied to reduce the feature space while preserving the most important patterns and trends in the data.

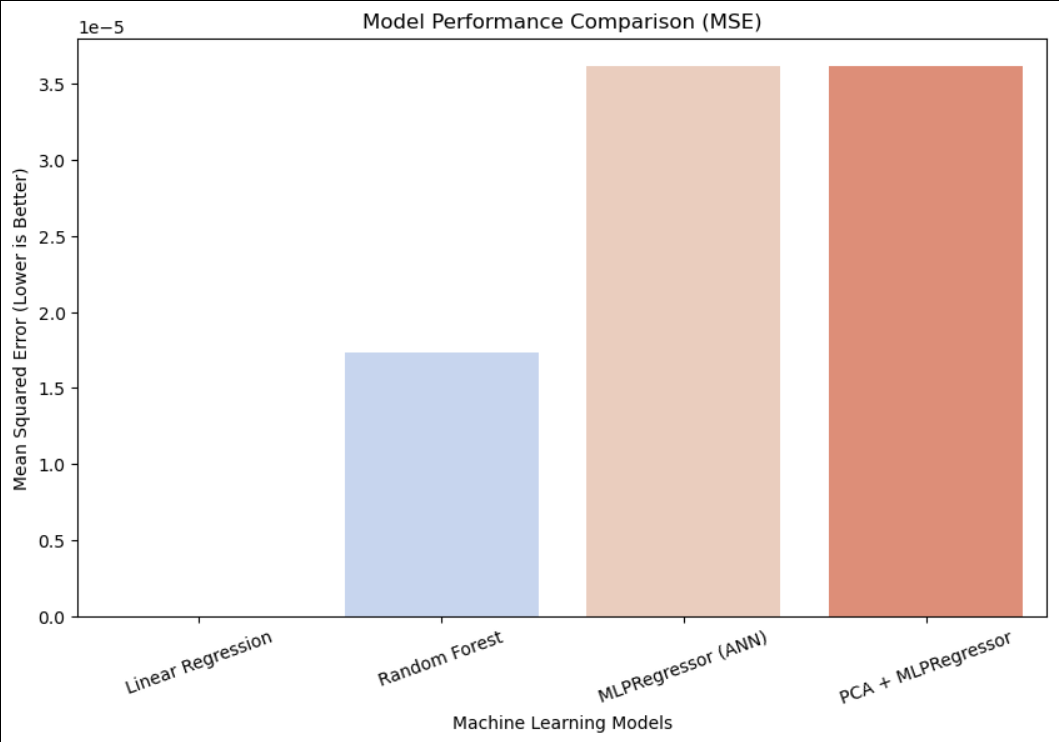


Fig 7. Model Performance Comparison (MSE)

**Chapter 6**

**Conclusion**

This project successfully applied multiple machine learning techniques—including Linear Regression, Random Forest, ANN, PCA, and LDA—to analyze and predict the progression of Covid-19 using historical data. Linear Regression offered a simple, interpretable approach for forecasting trends over time, acting as a baseline for evaluating more sophisticated models. Random Forest enhanced predictive performance by handling nonlinearities and feature interactions. The use of ANN (Artificial Neural Network) allowed the project to capture complex relationships in the data, and the integration of PCA helped in reducing noise and improving training efficiency. LDA contributed valuable insights into class-based analysis, especially in identifying patterns of high-risk or low-risk periods.

**Chapter 7**

**Future Work**

**Real-Time Data Integration:**

One major improvement would be integrating real-time or streaming Covid-19 data using APIs from trusted sources such as WHO or Johns Hopkins University. This would allow the models to continuously update and adapt to the latest trends, making them more relevant for real-time forecasting and early warning systems.

**More Advanced Deep Learning Models**

Although an ANN was implemented, deeper architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), especially LSTM (Long Short-Term Memory) networks, could be explored to better capture temporal dependencies and long-term patterns in time-series data.

**Chapter 8**

**GitHub Repository Link**

1. Student 1 Name : Kush Chauhan

Link-: <https://github.com/kush-prog/COVID-19-Data-Analysis-and-Prediction.git>

2. Student 2 Name : Kanak Jain

Link: <https://github.com/Kanakjain04>

1. Student 3 Name : Varun Kumar

Link: <https://github.com/varunkumarparawa>

1. Student 4 Name : Tanya Chaudhary

Link: <https://github.com/tanyachgithub>

**REFERENCES**

**[1]Kaggle** - Dataset for Covid-19 trend analysis and prediction:  
*Dataset Source*: [Kaggle - COVID-19 Data](https://www.kaggle.com/).

**[2**] **ChatGPT** - Assisted in generating the Python code documentation and crafting the report.

**[3**]**John Hopkins University COVID-19 Data Repository**:  
Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5), 533–534.  
[GitHub Repository](https://github.com/CSSEGISandData/COVID-19).

**[4] World Health Organization (WHO) – Coronavirus Dashboard**:  
World Health Organization. (2020). WHO Coronavirus (COVID-19) Dashboard.  
[WHO Dashboard](https://covid19.who.int/).

**[3Geron, A. (2019): *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.**

**[4**] **Pedregosa, F., Varoquaux, G., et al. (2011)**:

Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.  
[Research Paper](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html).

**[5**] **Bishop, C. M. (2006):*Pattern Recognition and Machine Learning*. Springer.**

**[6]Jolliffe, I. T., & Cadima, J. (2016):**

**Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).**[**DOI Link**](https://doi.org/10.1098/rsta.2015.0202)**.**