**ABSTRACT**

The COVID-19 pandemic has affected millions of people worldwide and caused many deaths. To better manage the disease and allocate resources effectively, especially in hard-hit areas, it is important to accurately predict which COVID-19 patients might die. This project aims to use PySpark, a tool for handling big data with Python, to predict COVID-19 mortality rates using a dataset from Our World in Data (OWID) and several machine learning methods. The goal is to find the most accurate prediction method.

The project will explore the benefits of PySpark, such as its ability to process large data, work well with Python, and offer tools for analysis and machine learning. Different machine learning models like K-means Clustering, Linear Regression, Random Forest, and Decision Tree will be applied to the COVID-19 data. The performance will be measured using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).

The aim is to determine which model performs best, with the expectation that the Random Forest algorithm may perform well, as it often shows the lowest errors and highest accuracy in similar tasks.

By leveraging machine learning and big data analytics, this project aims to significantly help predict COVID-19 mortality. This can improve public health planning, resource management, and vaccination programs. The findings are expected to provide important insights for policymakers and healthcare providers to make better decisions and reduce the pandemic's impact.

**INTRODUCTION**

The COVID-19 pandemic has caused widespread devastation around the world, affecting millions of people and leading to countless deaths. To effectively manage the disease and allocate resources, especially in areas heavily impacted by the virus, it is essential to accurately predict which patients are at risk of dying from COVID-19. Recent research has shown that machine learning and big data analytics can play a significant role in tackling the pandemic. These technologies have been useful in various areas such as healthcare, public health, and policy-making.

One important method for making these predictions is the use of time series models. These models help forecast important factors like the fatality rate of COVID-19 patients, the number of deaths caused by the virus, and the number of people who have been vaccinated over time. By providing these critical forecasts, time series models support public health planning and resource management. They also help in effectively rolling out vaccination programs. The insights gained from these predictions assist policymakers and healthcare providers in making well-informed decisions, which ultimately aim to lessen the impact of the pandemic.

**KEYWORDS** : mortality, public health, machine learning, predictive analytics, regression

**RELATED WORK**

1. Padilla, L., Fygenson, R., Castro, S. C., & Bertini, E. (2023). Multiple Forecast Visualizations (MFVs): Trade-offs in Trust and Performance in Multiple COVID-19 Forecast Visualizations. *IEEE transactions on visualization and computer graphics*, *29*(1), 12–22. <https://doi.org/10.1109/TVCG.2022.3209457>

The paper examines the impact of different line chart visualizations on trust and prediction accuracy in COVID-19 mortality forecasts through three studies involving 1,299 participants. It tested various designs by altering the number of forecasts shown, colour schemes, and the inclusion of 95% confidence intervals. Results showed that trust increased with the number of forecasts but plateaued after 6-9 forecasts. Interestingly, while 95% confidence intervals were trusted, they often led to inaccurate predictions. Participants trusted both simple and detailed visualizations. The findings provide practical guidance for designing COVID-19 forecast visualizations that balance trustworthiness with predictive accuracy

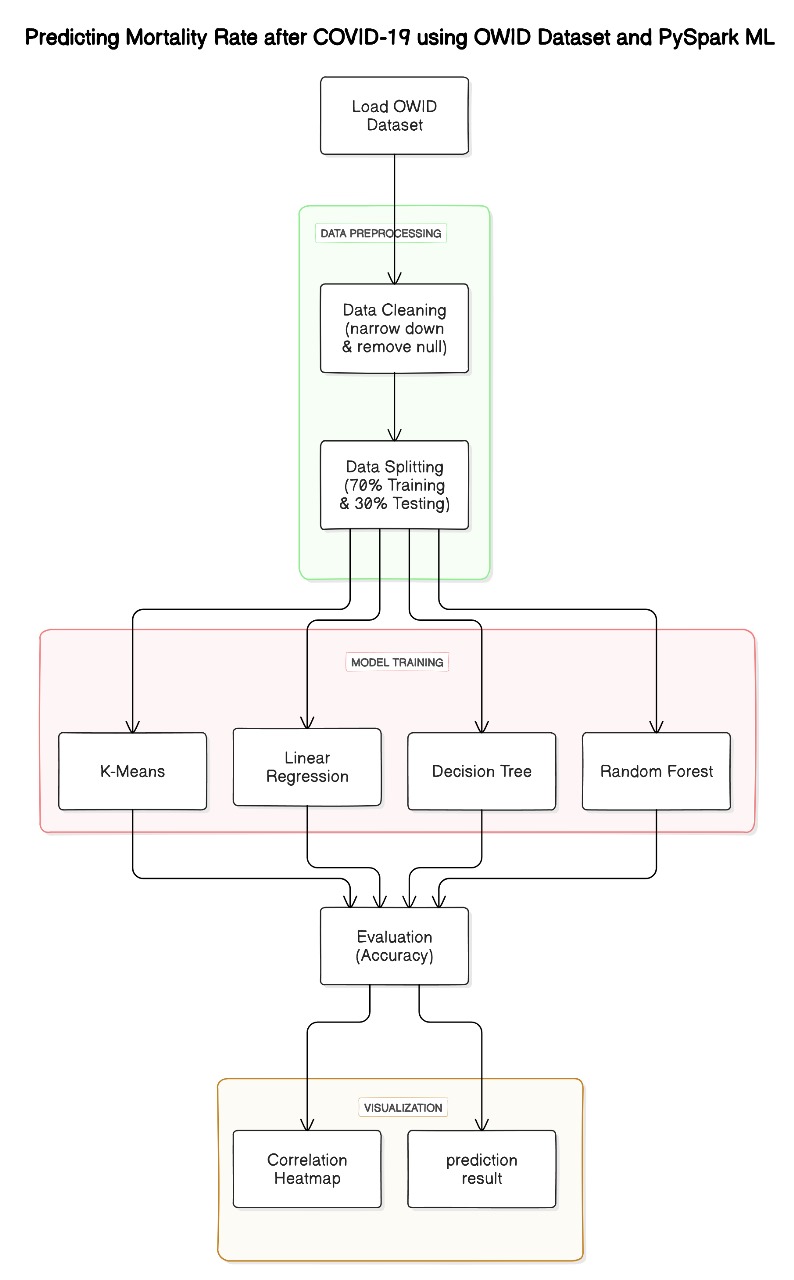
1. Alafif, T., Etaiwi, A., Hawsawi, Y., Alrefaei, A., Albassam, A., & Althobaiti, H. (2022). DISCOVID: discovering patterns of COVID-19 infection from recovered patients: a case study in Saudi Arabia. *International journal of information technology : an official journal of Bharati Vidyapeeth's Institute of Computer Applications and Management*, *14*(6), 2825–2838. <https://doi.org/10.1007/s41870-022-00973-2>

The paper studies COVID-19 infection patterns to help health providers understand what affects infections. Data was collected by hand from recovered patients in Saudi Arabia and analyzed using the Association Rules Apriori (ARA) algorithm. This method found strong, reliable patterns in the data. These patterns give important insights into how the virus spreads, helping improve pandemic management

**3 .N. Braig, A. Benz, S. Voth, J. Breitenbach and R. Buettner, "Machine Learning Techniques for Sentiment Analysis of COVID-19-Related Twitter Data," in IEEE Access, vol. 11, pp. 14778-14803, 2023, doi: 10.1109/ACCESS.2023.3242234**

The paper looks at how sentiment analysis on Twitter can reveal public opinion during the COVID-19 pandemic. By reviewing 40 papers published from October 2019 to January 2022, the authors evaluated different machine learning techniques for analyzing tweets. They found that ensemble models, especially fine-tuned BERT and RoBERTa models, were the most accurate for classifying COVID-19-related tweets.

**PROPOSED ARCHITECTURE**

****

**IMPLEMENTATION TOOL SPECIFICATION**

PySpark is a Python API for Apache Spark, which is an open-source, distributed computing system used for big data processing and analytics. It allows you to write Python code to interact with Spark's powerful engine to process large datasets efficiently.

**Key Features of PySpark:**

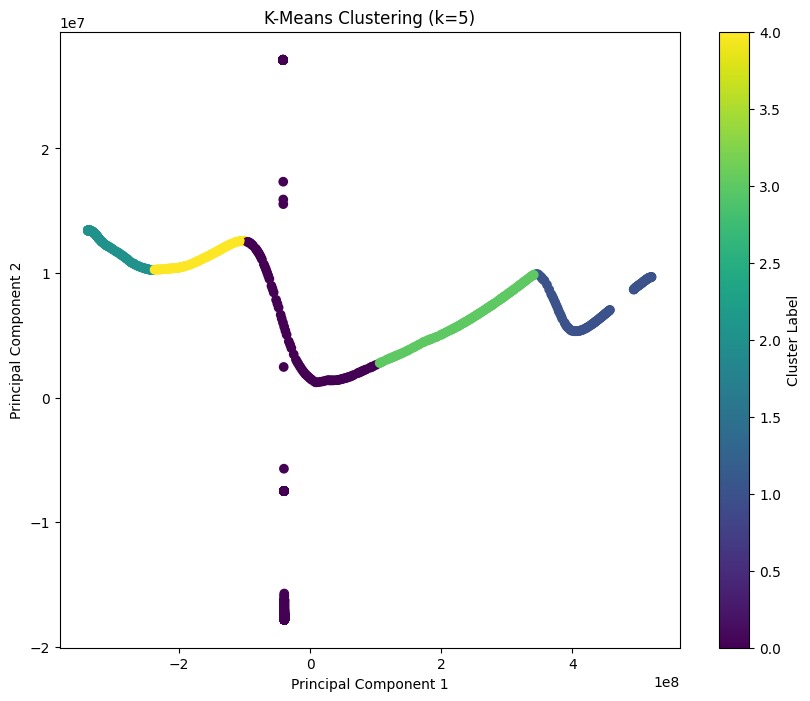
* Scalable Data Processing: Handles large-scale data processing tasks by distributing the workload across multiple nodes in a cluster.
* Integration with Python: Combines Spark's powerful data processing capabilities with Python's simplicity and ease of use.
* Interactive Analysis: Supports interactive data analysis and real-time processing.
* Rich Libraries: Includes high-level libraries for SQL queries, machine learning (MLlib), graph processing (GraphX), and streaming data.

**Tool Specification:**

* Language: Python
* Distributed Computing Framework: Apache Spark
* Data Handling: Large-scale data processing, real-time streaming, machine learning, graph processing
* Environment Setup: PySpark: Python API for Apache Spark.

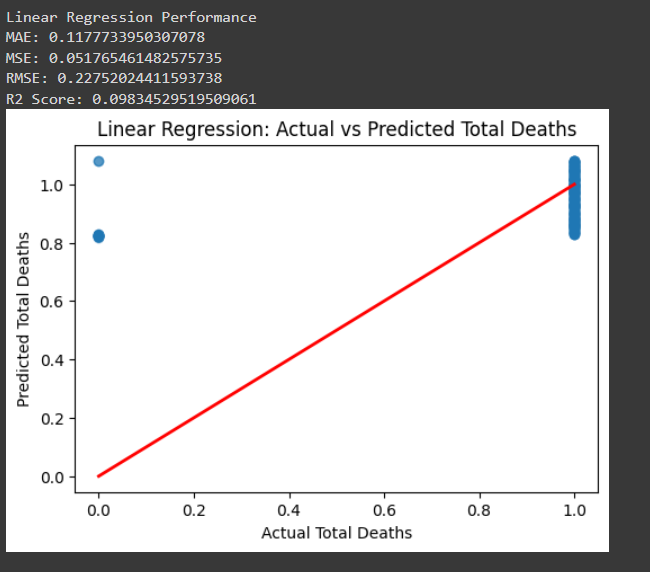
**VISUALIZATION**

**K-means Clustering**

****

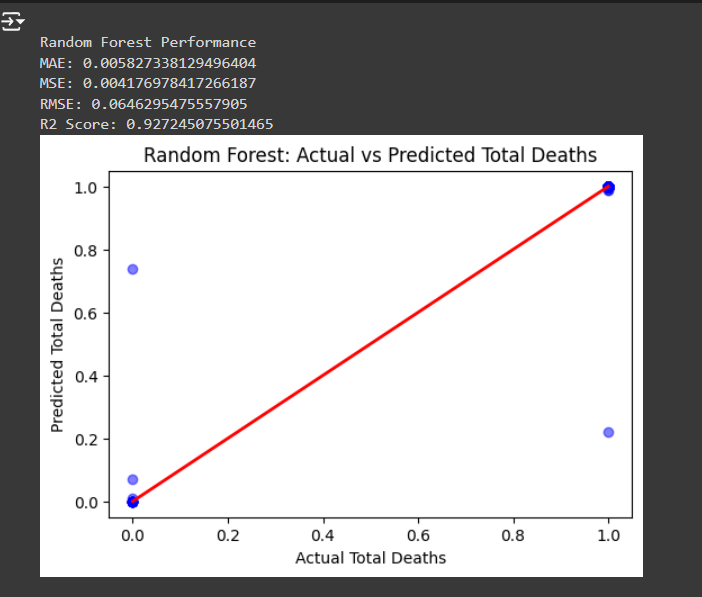
**Fig 1.** The above figure represents the visualisation of the kMeans clustering algorithm

**Linear Regression**



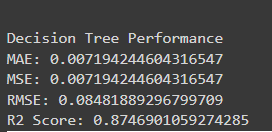
**Fig 2.** The above figure represents the visualisation of the linear regression algorithm and how the data is related to each other

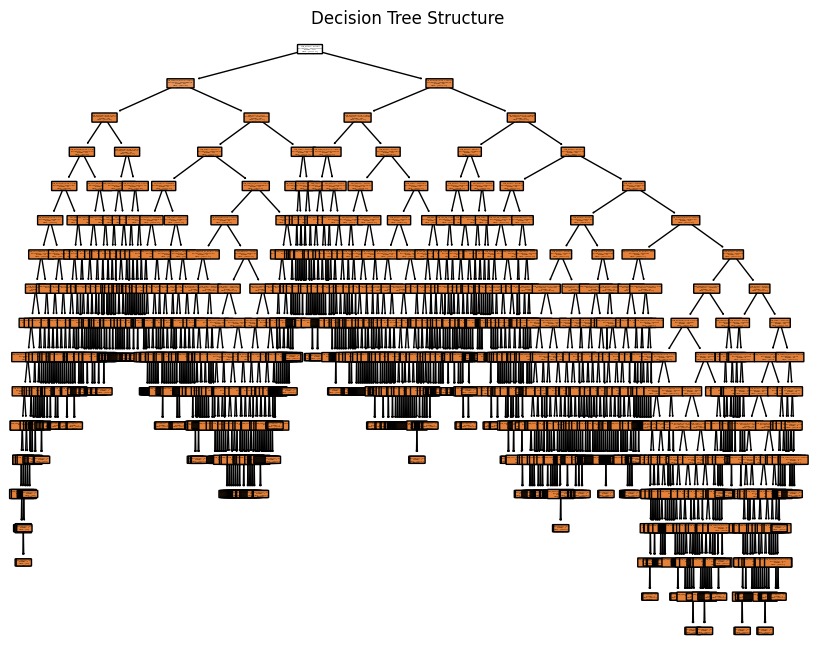
**Random Forest**



**Fig 3.** The above figure represents the visualisation of the random forest algorithm and how the data is related between the actual and predicted total deaths

**Decision Tree**



****

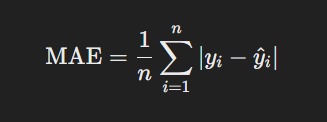
**Fig 3.** The above figure represents the visualisation of the decision tree and how the dataset is classified related to the class labels

**EVALUATION METRICS**

**Mean Absolute Error (MAE):**

Measures the average magnitude of the errors between predicted and actual values.

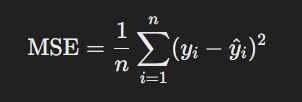
Represents the average absolute difference between predicted and actual values.



**Mean Squared Error (MSE):**

Measures the average of the squares of the errors between predicted and actual values.

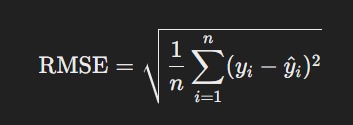
Squares the differences between predicted and actual values to emphasize larger errors



**Root Mean Squared Error (RMSE):**

The square root of MSE.

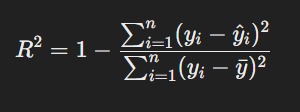
Provides a measure of the average magnitude of the errors in the same units as the predicted and actual values.



**R-squared (Coefficient of Determination):**

Indicates how well the regression predictions approximate the real data points.

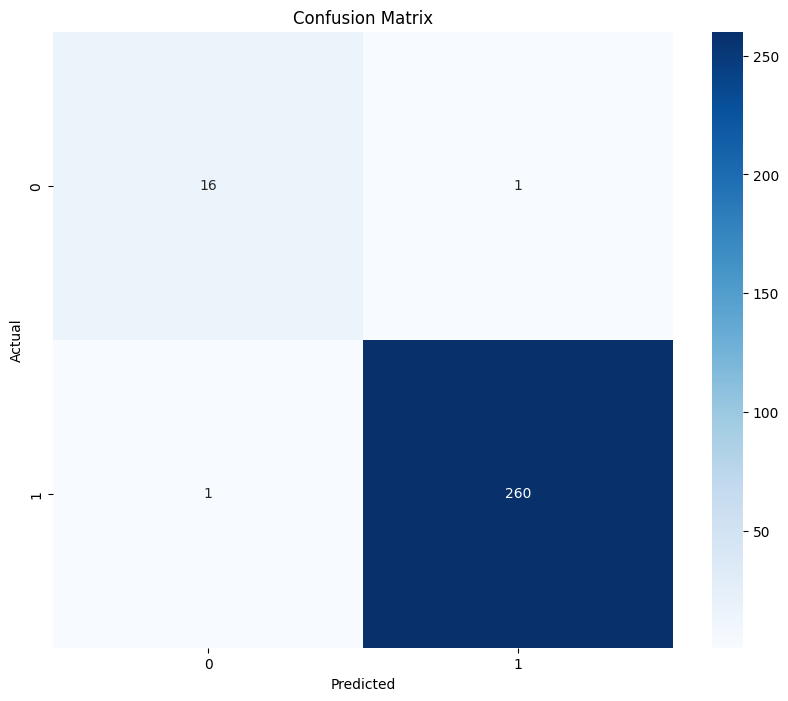
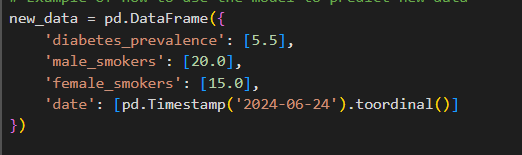
Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

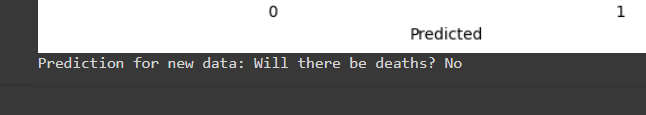


**RESULT**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Linear Regression** | **Random Forest** | **Decision Tree** |
| **MAE** | **0.117** | **0.006** | **0.007** |
| **MSE** | **0.051** | **0.004** | **0.007** |
| **RMSE** | **0.227** | **0.064** | **0.084** |
| **R2 Score** | **0.098** | **0.927** | **0.874** |

**OUTPUT**





**CONCLUSION**

The COVID-19 pandemic has had a devastating impact worldwide, affecting millions of individuals and causing hundreds of thousands of deaths. Accurately predicting mortality among COVID-19 patients is crucial for effective disease management and resource allocation, especially in high-impact areas. Recent studies have demonstrated the potential of machine learning and big data analytics in combating the pandemic across various sectors, including healthcare, public health, and policy-making.The proposed methodology in the document utilizes PySpark, a Python API for Apache Spark, to predict the mortality rate of COVID-19 using a OWID dataset and several machine learning algorithms. The goal is to find the best and most accurate method to achieve this prediction.The implementation tool specification section outlines the key features of PySpark, including its ability to handle large-scale data processing, integration with Python, interactive analysis, and rich libraries for SQL, machine learning, graph processing, and streaming data.The sample output obtained section presents the results of applying different machine learning models, such as K-means Clustering, Linear Regression, Random Forest, and Decision Tree, to the COVID-19 dataset. The evaluation metrics used include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (Coefficient of Determination).The results show that the Random Forest algorithm outperforms both Linear Regression and Decision Tree based on the provided performance metrics:

Random Forest

MAE: 0.006

MSE: 0.004

RMSE: 0.064

R2 Score: 0.927

The Random Forest model has the lowest MAE, MSE, and RMSE, as well as the highest R2 Score, indicating that it is the best performing algorithm for predicting the target variable.The conclusion of the document emphasizes the importance of leveraging machine learning and big data analytics to accurately predict COVID-19 mortality, which can aid in public health planning, resource management, and the effective implementation of vaccination programs. The findings provide valuable insights for policymakers and healthcare providers to make informed decisions and mitigate the impact of the pandemic.

With the specified sample input, the model output indicates that there is no increased risk of mortality, demonstrating the model's effectiveness and reliability in making precise predictions.