**Predictive Modelling of COVID-19 Metrics**

**Introduction**   
The COVID-19 pandemic has significantly influenced global health, economies, and societies, emphasizing the need for data-driven strategies to understand and mitigate its impact. Predictive modelling is essential for identifying trends, forecasting outcomes, and guiding public health policies. This project employs data science techniques to analyse COVID-19 data, develop predictive models, and extract actionable insights. Using exploratory data analysis (EDA) and linear regression, the relationships among confirmed cases, recoveries, active cases, and deaths were examined to forecast pandemic outcomes and inform decision-making.

The study utilized a dataset from Kaggle's CORD-19 repository, which includes country-specific COVID-19 metrics. Key objectives were to clean and enhance the dataset, identify trends and anomalies through EDA, apply machine learning techniques for predictions, and interpret results to provide valuable public health insights.

**Data Preparation**  
Robust data preparation is foundational to effective predictive modelling. The raw COVID-19 dataset underwent several pre-processing steps to ensure its suitability for analysis:

1. **Data Cleaning:**
   * Missing values were handled through imputation or removal to maintain data integrity.
   * Duplicate entries were eliminated to ensure consistency.
2. **Data Transformation:**
   * Percentage-based variables, such as Deaths per 100 Cases and Recovered per 100 Cases, were converted into numerical formats.
   * Metrics normalized to population size, such as cases per 100,000 people, were calculated to allow fair regional comparisons.
3. **Feature Engineering:**
   * New variables, such as growth rates and mortality ratios, were created to better capture pandemic dynamics.

These reprocessing steps established a reliable dataset, enabling insightful analysis and predictive modelling.

**Exploratory Data Analysis(EDA)**  
EDA was instrumental in uncovering patterns and relationships within the dataset. Notable insights included:

1. **Trends in Case Counts:**
   * Bar charts highlighted the top 10 countries with the highest confirmed cases, revealing significant regional disparities.
   * Time-series plots illustrated case number trends over time, identifying peaks and patterns.
2. **Mortality and Recovery Rates:**
   * Metrics such as Deaths per 100 Cases and Recovered per 100 Cases revealed regions with high mortality rates and varied recovery outcomes.
   * Box plots comparing recovery rates across WHO regions provided insights into healthcare system effectiveness.
3. **Correlation Analysis:**
   * A heat map of numerical variables (e.g., confirmed cases, recoveries, deaths) identified strong correlations, particularly between confirmed cases and deaths.
   * These findings informed the selection of input variables for predictive modelling.

The EDA phase provided a comprehensive understanding of COVID-19's progression and impact, laying the groundwork for model development.

**Model Development**  
Linear Regression was employed to forecast COVID-19 deaths. Key steps included:

1. **Variable Selection:**
   * Independent variables (Confirmed Cases, Recovered Cases, and Active Cases) were chosen based on their relevance and correlation with the target variable (Deaths).
2. **Model Training and Testing:**
   * The dataset was divided into training (80%) and testing (20%) subsets to ensure model generalizability.
   * A Linear Regression model was trained on the training data and evaluated on the test data.
3. **Model Evaluation:**
   * Performance metrics such as Mean Squared Error (MSE) and R-squared were calculated. An R-squared score of 1.00 indicated an excellent fit, while a low MSE validated the model’s accuracy.
   * A scatter plot comparing actual and predicted deaths demonstrated high model precision, with minimal deviations.

The model effectively captured the relationships among variables, highlighting the potential of predictive techniques in understanding pandemic trends.

**Results and Insight**

Key findings from the analysis include:

1. **Model Performance:**
   * The Linear Regression model achieved an R-squared score of 1.00, explaining nearly all variability in deaths.
2. **Feature Impact:**
   * Confirmed Cases emerged as the most critical predictor, followed by Active Cases and Recovered Cases.
3. **Practical Implications:**
   * Regions with high death rates but low recovery rates were identified as priorities for resource allocation and healthcare system improvements.
   * Insights into recovery rates underscored the importance of timely medical interventions in reducing mortality.
4. **Visual Interpretations:**
   * Heat maps and regional recovery rate comparisons provided actionable insights for policymakers on addressing disparities and enhancing interventions.

**Conclusion**   
This project underscores the importance of data-driven approaches in managing public health crises. By combining EDA and predictive modelling, critical insights into COVID-19 metrics were generated, providing a solid foundation for informed decision-making. The strong performance of the Linear Regression model highlights the significance of thorough data preparation, variable selection, and evaluation. Future work could extend these models by incorporating additional variables and advanced techniques to further support global health initiatives.