3MTT CAPSTONE PROJECT TECHNICAL REPORT

BY Azeez Olabode Adewale

INTRODUCTION

The objective is to explore the data, developed predictive models and provide insight of the spread of Covid-19

DATA COLLECTION

The dataset was sourced from Kaggle .

DATA PREPROCESSING

I imported the necessary libraries for the analysis, then loading the data from the file and examining the dataset. This allows me to determine the necessary cleaning steps and identify column needed for my analysis.

PREVIEWING THE DATA

The dataset consists of 187 entries and 15 columns, capturing various Covid-19 metrics for different countries and regions. The summary of the column is listed below.

- Country/Region: The country or region name
- Confirm, Deaths Recovered, and Active: Case count by status
- New Cases, New deaths, New recovered: Recent case changes
- Deaths/100 Cases, Recovered/100 Cases, Death /100 Recovered: Ratio per percentage
- Confirmed last week, 1 week change, 1 week % increase: Weekly case changes
- WHO Region: Assigned region From the data, there are no missing values

EXPLORATORY DATA ANALYSIS

Univariate Analysis

- Hisplot and boxplot shows the distribution of the cases of Covid-19. There is a noticeable spread with the data with outliers
- Confirmed Cases and Deaths Distribution are rightly-skilled which indicate a few countries have high count compared with the majority
- Recovered and Active cases follow a similar distribution pattern.

Multivariate Analysis

Catplot shows the distribution of case count by status of Covid-19 across the country, from the plot it shows that;

- Few countries show significantly higher active cases but US has the highest number of active cases of Covid-19 while Saints Kitts, Mauritius, Western Sahara have relatively low active cases with Holy See, Sam Marino, Brunei, Domica and Grenada have no active cases
- For the confirmed cases, the US has the highest number follow by Brazil, India, and Russia while Greenland, Holy Sea and Western Sahara reported least number of confirmed cases.

- For the death cases, the US recorded the highest number, followed by Brazil, United Kingdom, Mexico, Italy, Indian, France and Spain with high cases of death while Seychelles, Laos and others recorded no cases of death.

  For the recovered, Brazil, US, India and Russia recorded the highest cases with Greenland Holy See, Papua New Guinea recorded least cases while Mozambique, Canada, Sweden and others recorded no cases.

  An additional exploratory Analysis was performed to identify the trends in Covid-19 cases as well as the correlation between different Metrics.
- Line plot was used to illustrate the distribution of Confirmed Death and Recovered Cases across different WHO region. The visualization helps to observe which region were most affected by Covid-19
- Americas recorded highest number of Confirmed and Death Cases
- Africa and Western Pacific have relatively Low Confirmed and Death cases

Correlation Metrics of Covid -19

- Strong positive correlation exists between confirmed and death (0.93), recovered (0.91) and active (0.93)
- New cases are confirmed 0.91 and new death (0.97)
- Percentages increase in cases and death shows weak correlation with the total count suggesting other factor may influence weekly growth rate.

MODEL DEVELOPMENT

Model Selection

Linear regression model was trained for predicting active cases based on other feature such as confirmed, death recovered new cases, new death and new recovered. The features were selected because they correlate with active cases.

Trained-test Split

80% was used to train while 20% to test.

Model Training

The linear regression model was trained to learn the relationship between the features and the number of active cases

MODEL EVALUATION

The model performance was evaluated using Root Mean Square Error (RMSE) and R-Square.

The RMSE shows the average difference between predicted and active cases meaning that a lower RMSE indicates better performance.

Also the R-square shows how well the model explain the variance in the data with value closer to 1 indicating a better fit

- From the Model, RMSE is approximately 2.70 e^{-11} which means the model is performing well.
- R-Square is equal to 1 which means a better fit i.e the model is performing 100% which could be due to data leakage.

Coefficient and Intercept

The linear regression model provides coefficient for each feature, the expected changes in active cases for a 1 unit change in the feature if other features are kept constant.

From the coefficient, it can be deduced that

- Confirmed cases, Deaths and Recovered are the major features contributing to the active cases of Covid-19
- New cases, new death and new recovered have no or little influence on active cases of Covid19

INSIGHT

The predictive model provides actionable insight for managing active Covid-19 cases effectively by implementing timely in regions at the risk of becoming overwhelmed by high active case count.

CONCLUSSION

This model helps in understanding the factor that influence active cases and this can be used to forecast active cases based on current data, for instance knowing how confirmed cases and recovered and death impact active cases allows health organizations to prepare resources based on anticipated active cases supporting effective resources planning