Technical Report on Predictive Modelling for COVID-19 in Public Health

1. Introduction

The COVID-19 pandemic has presented unprecedented challenges for public health organizations worldwide. Accurate prediction of the virus's spread and understanding the factors that influence transmission and patient outcomes are critical for effective policy-making and resource allocation. This report details the development of a predictive modeling system for HealthGuard Analytics, aimed at generating actionable insights from historical COVID-19 data. The project encompasses data cleaning, exploratory data analysis (EDA), predictive modeling, and the presentation of findings and recommendations for public health responses.

2. Methodology

2.1 Data Collection

The primary dataset utilized for this analysis is the COVID-19 Open Research Dataset (CORD-19), which is publicly available on Kaggle. This dataset includes comprehensive information on COVID-19 case counts, demographic data, and various health metrics essential for understanding the pandemic's dynamics.

Dataset Source: CORD-19 on Kaggle

2.2 Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and integrity of the dataset for modeling. The following procedures were implemented:

- **Cleaning:**
  - Missing values were addressed using deletion and imputation techniques.
  - Duplicate entries were removed to maintain the uniqueness of records.
  - Date and location formats were standardized for consistency.
- **Transformation:**
  - Numerical features were normalized to ensure comparability across different scales.

- **Feature Engineering:**
    - Derived variables were created, such as daily growth rates, mortality ratios, and cases per population, to enhance the dataset's analytical depth.

2.3 Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns, correlations, and anomalies within the data. Key activities included:

- **Visualizations:**
    - Various charts were employed, including line plots, bar charts, and scatter plots, to visualize trends in COVID-19 cases, mortality rates, and recoveries over time.
- **Key Insights:**
    - Identification of demographic and environmental factors influencing the spread and severity of COVID-19 cases.

2.4 Model Development

Predictive modeling was conducted using both time-series and classification approaches to forecast COVID-19 trends:

- **Machine Learning Models:**
    - Linear Regression was employed to predict the number of deaths based on confirmed cases and recoveries.
- **Evaluation:**
    - Model performance was assessed using metrics such as Mean Squared Error (MSE) and R-squared score to evaluate prediction accuracy.

3. Results

3.1 Data Cleaning and Preparation

The initial dataset was loaded and examined. Following data cleaning, a summary of the cleaned dataset is presented below:

```
import pandas as pd

df = pd.read_csv("assets/data.csv")
```

```
print("Dataset Head:", df.head())
print("\nDataset Info:", df.info())

# Clean the data
df_cleaned = df.copy()
df_cleaned = df_cleaned.dropna()
df_cleaned = df_cleaned.drop_duplicates()
df_cleaned = df_cleaned.reset_index(drop=True)
```

3.2 Exploratory Data Analysis

Several visualizations were generated to illustrate key findings:

1. **Top 10 Countries by Confirmed Cases:**

```
plt.figure(figsize=(12, 6))
top_10_confirmed = df_cleaned.nlargest(10, 'Confirmed')
sns.barplot(data=top_10_confirmed, x='Country/Region', y='Confirmed')
plt.xticks(rotation=45)
plt.title('Top 10 Countries by Confirmed Cases')
plt.show()
```

2. **Death Rate Analysis:**

```
plt.figure(figsize=(12, 6))
top_10_death_rate = df_cleaned.nlargest(10, 'Deaths / 100 Cases')
sns.barplot(data=top_10_death_rate, x='Country/Region', y='Deaths /
100 Cases')
plt.xticks(rotation=45)
plt.title('Top 10 Countries by Death Rate')
plt.show()
```

3. **Recovery Rate by WHO Region:**

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_cleaned, x='WHO Region', y='Recovered / 100
Cases')
plt.xticks(rotation=45)
```

```
plt.title('Recovery Rate Distribution by WHO Region')
plt.show()
```

4. **Correlation Matrix of COVID-19 Metrics:**

```
numeric_cols = ['Confirmed', 'Deaths', 'Recovered', 'Active', 'Deaths
/ 100 Cases', 'Recovered / 100 Cases']
correlation_matrix = df_cleaned[numeric_cols].corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of COVID-19 Metrics')
plt.show()
```

3.3 Model Development

The modeling phase involved training a Linear Regression model to predict the number of deaths based on confirmed and recovered cases.

```
X = df_cleaned[['Confirmed', 'Recovered', 'Active']]
y = df_cleaned['Deaths']
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Train linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)
```

3.4 Model Evaluation

The model's performance was evaluated using MSE and R-squared metrics:

```
from sklearn.metrics import mean_squared_error

mse = mean_squared_error(y_test, y_pred)
```

```
r2 = model.score(X_test, y_test)

print(f'Mean Squared Error: {mse:.2f}')
print(f'R-squared Score: {r2:.2f}')
```

Predictions vs Actual Deaths

```
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
'r--', lw=2)
plt.xlabel('Actual Deaths')
plt.ylabel('Predicted Deaths')
plt.title('Model Predictions vs Actual Deaths')
plt.show()
```

3.5 Feature Importance

The importance of each feature in the model was assessed:

```
feature_importance = pd.DataFrame({'Feature': X.columns,'Coefficient':
model.coef_})
print("\nFeature Importance:")
print(feature_importance)
```

4. Conclusion and Recommendations

The predictive modeling system developed for HealthGuard Analytics successfully provides insights into the dynamics of COVID-19. The findings indicate that confirmed cases and recoveries are significant predictors of death rates, highlighting the importance of timely intervention in controlling outbreaks.

Recommendations for Public Health Responses:

1. **Resource Allocation:** Prioritize resources in regions with high confirmed case counts and low recovery rates.
2. **Policy Making:** Use predictive insights to inform policies aimed at controlling the spread of the virus, such as lockdowns and vaccination campaigns.
3. **Continuous Monitoring:** Implement continuous data collection and model updates to refine predictions and adapt to emerging trends.

5. Deliverables

- **Technical Report:** A detailed report covering all phases of the project.
- **Code Repository:** A well-documented GitHub repository containing all code used in the project.
- **Presentation:** A slide deck summarizing key findings and public health implications.

References

1. COVID-19 Open Research Dataset (CORD-19). Kaggle. Link
2. Python libraries: Pandas, Seaborn, Matplotlib, Scikit-learn for data analysis and visualization.