Kevin Muraga Njuguna SCT221-0235/2021

George Kibarua SCT221-0521/2021

Zachary Kinagwi SCT221-0229/2021

James nyoro SCT221-0086/2021

Ann Mutua SCT221-0177/2021

Data Collection:

COVID-19 Data Sources:

JHU, ECDC, Worldometer, ArcGIS, Africa Open Data, Kenya Open Data.

Gather data related to COVID-19 cases, deaths, recoveries, demographics, etc., focusing on Kenya-specific information.

2. Ingestion into Hadoop DFS Data Lake:

Transfer the collected data into Hadoop Distributed File System (HDFS).

3. Data Extraction using PySpark:

Utilize PySpark for reading and extracting data from the Hadoop Data Lake.

Consider the schema, data types, and structure of the extracted data.

4. Data Preprocessing:

Data Cleaning:

Handle missing values, duplicates, inconsistencies, etc.

Feature Engineering:

Create new features if necessary, e.g., calculate mortality rate, recovery rate, etc.

Normalization/Standardization:

Scale numerical features if needed.

Feature Selection:

Identify relevant features for modeling.

5. Predictive Analytics Techniques:

Model Selection:

Choose appropriate algorithms (Regression, Time Series, etc.) for prediction.

Consider techniques like Linear Regression, Decision Trees, Random Forest, LSTM (for time-series data), etc.

Train/Test Split:

Split the data into training and testing sets.

6. Model Building:

Feature Encoding:

Encode categorical variables if required.

Model Training:

Train the chosen predictive model using the training data.

Model Evaluation:

Evaluate the model's performance using appropriate metrics (e.g., RMSE, MAE, R-squared for regression).

7. Visualization:

Data Visualization:

Use libraries like Matplotlib, Seaborn, or Plotly to visualize the data.

Plot trends, patterns, and predictions.

8. Model Testing:

Testing:

Use the test dataset to predict outcomes.

Evaluate model performance on unseen data.

Iterate and Improve:

Fine-tune parameters, consider feature selection, or try different models if needed.

Important Considerations:

Resource Allocation: Ensure Hadoop cluster resources are adequate for processing large-scale data.

Data Security and Privacy: Handle sensitive data securely.
Data Compilation

(i) Description of Data Compilation

Explain the process used to compile the data.

Include code snippets/screenshots of the code used and the corresponding output.

Task 2: Data Ingestion into Hadoop Data Lake

(ii) Description of Data Ingestion

Describe the process of ingesting data into the Hadoop data lake.

Include relevant screenshots demonstrating this process.

Task 3: Data Extraction using PySpark

(iii) Description of Data Extraction

Detail how the data was extracted using PySpark.

Include screenshots showcasing the steps involved in data extraction.

Task 4: Data Pre-processing

(iv) Description of Pre-processing Tasks

Explain the pre-processing techniques utilized to prepare the data.

Include screenshots illustrating the pre-processing steps taken.

Justify the chosen techniques with reasons (e.g., data quality improvement, normalization, etc.).

Task 5: Test Results and Interpretations

(v) Test Results

Present the results obtained from the processed data.

Interpret the outcomes of the tests performed.

Task 6: Validation Results

(vi) Validation Results

Show the validation outcomes based on the processed data.

Interpret the validation results obtained.

Task 7: Potential Applications

(vii) Potential Applications

Discuss potential applications of the interpreted results.

Explain how these results could be used or applied in practical scenarios.