Big Data and Cloud Computing

CMP 4011

Phase 1

Team 5		
Name	**Section**	**BN**
Andrew Boshra	**1**	**17**
Peter Micheal	**1**	**21**
Taher Mohamed Ahmed	**1**	**37**
Omar Mohamed Ahmed	**2**	**6**

Problem statement

The outbreak of COVID-19 pandemic has created a global health crisis that has affected millions of people worldwide. This project aims to utilize a dataset on coronavirus to classify patients based on their symptoms and study the association rules between the symptoms and other routines and diseases they have. The main objective is to develop a predictive model that can accurately identify the patients at risk of contracting the virus and provide early interventions to prevent its spread. Additionally, the study of association rules will help in identifying the co-occurrence of symptoms and other underlying diseases, which will aid in the early diagnosis and management of the disease.

Dataset

Link	https://www.kaggle.com/datasets/meirnizri/covid19-dataset
Number of features	21 unique features
Number of records	1,048,576 unique patients
Size	58.5 MB
Other notes	The dataset was provided by the Mexican government

Planned approach

To preprocess the data, we will use multiple techniques, including **MapReduce**, to convert COVID diagnosis scores of 1-3 to 1 and scores of 4-7 to 0, along with other necessary prepossessing steps. We will leverage the power of **Spark**, **Spark SQL**, and **Spark MLlib** technologies to execute comprehensive data processing and analysis. For patient **classification**, we will use **logistic regression** to determine their COVID status based on symptom analysis. We will also employ the **Apriori Algorithm** to identify **association rules** between patients' symptoms and other routines and diseases, contributing to better disease diagnosis and management.