Fayoum University
Faculty of Computers and Information
ComputerScience Department

STREAMING DATA ANALYTICS USING MACHINE LEARNING ON LARGE SCALE SYSTEMS

Submitted By:

Fawzya Ramadan Sayed Hassan

Lecturer Assistant, Computer Science Department,
Faculty of Computers and Information, Fayoum University

A Thesis Submitted to the Faculty of Computers and Information, Fayoum University

In Partial Fulfillment of the Requirements for the Degree of Ph.D. in Computer Science

Prof. Abdelmgeid Amin Ali

Computer Science Department
Faculty of computers and information
Mina University

Dr. Masoud Esmail Masoud Shaheen

Computer Science Department
Faculty of computers and information
Fayoum University

SupervisedBy:

Faculty of Computer and Information

Fayoum University

2020

Abstract

Streaming Data Analytics in healthcare becomes a promising research direction due to the popularity of the real-time monitoring and tracking systems. Due to the enormous amount of healthcare streaming data and its higher speed, it is difficult to ingest, process, and analyze such huge data to make real-time actions in case of emergencies by using traditional methods. Therefore, the work in this dissertation concerns about how to build a real-time system that can handle streaming data from health-based social streaming data or wearable medical sensors and indicate the current status for the patient health. This has been done by introducing two systems that considered real-time data to improve streaming data analytics.

The first contribution is called The Real-time Diabetes Disease Prediction System. It is developed to predict diabetes disease from health-based social streaming data to indicate patient health status. The proposed system aims to find the most accurate machine learning modelwhich has the highest accuracy of diabetes prediction. The experimental results have determined that the Random Forest (RF) model has achieved the highest accuracy among other models at 84.11%. For online prediction through social media, the system handled streaming Twitter data about patients' health. In doing so, Kafka and Spark streaming are integrated into the backend of the proposed system. Then, the FR classifier is used to predict the patient's current health status in real-time.

The second contribution is called the Online Prediction System. The proposed system focuses on applying streaming machine learning models on streaming health data events ingested to spark streaming through Kafka topics. The experimental results are done on the historical medical datasets and simulated wearable medical sensor data. The experimental results have

proved that the online prediction system can online learn and update the model according to the new data arrival and window size.