Big Data Analytics in Healthcare: Case Study - Miscarriage Prediction

Hiba Asri, OSER Laboratory, Faculty of Sciences and Technologies, Cadi Ayyad University, Marrakesh, Morocco
Hajar Mousannif, LISI Laboratory, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakesh, Morocco
Hassan Al Moatassime, OSER Laboratory, Faculty of Sciences and Technologies, Cadi Ayyad University, Marrakesh, Morocco

ABSTRACT

Sensors and mobile phones shine in the Big Data area due to their capabilities to retrieve a huge amount of real-time data; which was not possible previously. In the specific field of healthcare, we can now collect data related to human behavior and lifestyle for better understanding. This pushed us to benefit from such technologies for early miscarriage prediction. This research study proposes to combine the use of Big Data analytics and data mining models applied to smartphones real-time generated data. A K-means data mining algorithm is used for clustering the dataset and results are transmitted to pregnant woman to make quick decisions; with the intervention of her doctor; through an android mobile application that we created. As well, she receives recommendations based on her behavior. We used real-world data to validate the system and assess its performance and effectiveness. Experiments were made using the Big Data Platform Databricks.

KEYWORDS

Big Data, Cluster, Data Mining, Databricks, Kmeans, Miscarriage Prediction, Spark

INTRODUCTION

Miscarriage is the spontaneous loss of a pregnancy before the 20th week and presents the most common hurtful pregnancy outcome. As miscarriage is an irreversible phenomenon, prediction and prevention are still the only way to deal with this problem, which has negative psychological consequences for the woman and her partner as well (Kong et al., 2010). Some studies report that 50 percent of pregnancies will end in miscarriage (Arck et al., 2008).

Well-established risk factors for miscarriage include history of miscarriage, increased maternal age and infertility (Khalil et al., 2013; Waldron, n.d.). Also, several behavioral and social risks contribute in the increasing of the risk of miscarriage. Smoking (Pineles, Park, and Samet, 2014), alcohol drinking and caffeine intake (Avalos et al. 2014; Chen et al., 2016) are the main examples, but they still unconfirmed because some studies show these associations in the context of nausea which reduce the risk of miscarriage (Wen et al., 2001).
Stress and emotional wellbeing remain also important in pregnancy. In fact, the combination between life events and stress degree of the person increases risk of getting miscarriage (Khan et al., 2017). Emotions like anxiety, shock, sadness, anger, blame, depression, sleep disturbance have all been described as emotional responses to pregnancy loss (Mills et al., 2014). On other aspect, women are asked to participate in regular and moderate exercises while pregnant. Physical activity is defined by using either individual characteristics of a woman’s occupation such as standing, stooping, lifting, or walking. However, extreme or strong activity of the body has been associated with an elevated risk of miscarriage (Wong et al., 2010).

There is also increasing interest in the role that eating well plays in pregnancy (Women, 2005). Food safety is important for everyone and especially for pregnant women, because of her susceptibility to getting an illness during pregnancy. Eating in a restaurant, snack or cafe presents a higher risk of developing certain foodborne illnesses to pregnant women and their unborn children, and that can conduct to a miscarriage. Obesity is also associated with an increased risk of pregnancy complications. Indeed, obesity represents an independent cause of miscarriage (Pasquali, 2006), with rates that vary between 17 and 27 percent.

It becomes clear that several risk factors contribute to miscarriages. The research problem this paper tries to address is how reality mining, Big Data analytics and machine learning algorithms can help pregnant women to avoid getting miscarriage.

The idea is to take advantage of the ease with which mobiles and sensors can capture daily data and apply data mining techniques and big data predictive analytics to this data to make pregnant women’s life safer and healthier. Thanks to sensors and mobile phones, we can collect data from citizens, systems, and general things that weren’t previously possible (Asri, Mousannif, and Al Moatassime, 2018; Wang et al., n.d.).

The rest of this paper is illustrated over a few sections starting with a brief literature review followed by an overview of miscarriage prediction and detailed prediction process in section 3. Section 4 presents the data conceptual modeling; while section 5 presents K-means algorithm as data mining tool for clustering data. Section 6 illustrates the environment of the experiment, metrics, and experimental results. Finally, we conclude the paper by a conclusion and future work.

**BACKGROUND**

In the past few years, many researches were centered on the use of data mining tools in predicting outcomes. In the specific field of healthcare, data mining shows its power in predicting disease, extract patterns and make decisions. Several healthcare solutions are presented in the markets, which take a value from big data to help people make decisions by themselves to save time (Asri et al., 2015).

Mallu and Ezhilarasie (2015) explained the effective and useful sides of big data tools and techniques in healthcare system. They affirm that the integration of big data tools, data mining, medical informatics and big data analytics techniques; is an effective way on healthcare delivery costing and good healthcare outcome.

Chakraborty, Nagwani, and Dey (2012) assert that clustering is the most powerful tool used in several forecasting domains. In their study, they propose a methodology of k-mean clustering for weather forecasting. They use the air pollution dataset of west Bengal. In fact, the clustering process is made by K-mean algorithm and they got a list of weather categories based on the peak mean values of the clusters. When new data are presenting, K-mean affects it to the appropriate defined cluster. Therefore, they were able to predict weather events. Performance and correctness of their approach is also measured.

A performance comparison of four main machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naïve Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted by Asri et al. (2016). The main goal of the study is to achieve the correctness in classifying data with respect to efficiency and effectiveness of each
Related Content

Agent-Based Infrastructure for Dynamic Composition of Grid Services
www.igi-global.com/chapter/agent-based-infrastructure-dynamic-composition/40833?camid=4v1a

Sustainable Consumerism via Context-Aware Shopping
www.igi-global.com/article/sustainable-consumerism-via-context-aware-shopping/188859?camid=4v1a
Distributed Image Processing on a Blackboard System
www.igi-global.com/chapter/distributed-image-processing-blackboard-system/20523?camid=4v1a

Applications of Wireless Sensor Networks
www.igi-global.com/chapter/applications-wireless-sensor-networks/40840?camid=4v1a