Minor Project- Report

Aug-2021-2022

Course Faculty: Dr. Vinothini C Semester: 7

Course Name & code: BIG DATA ANALYTICS (19CS7DCBDA) Date:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TITLE OF THE PROJECT | Stroke Analysis Using Pyspark | | | |
|  |  | | | |
| STUDENT NAME | Aryan Khandelwal | Ayush Shah | Bhumika R |  |
| USN | 1DS19CS031 | 1DS19CS033 | 1DS19CS039 |  |
| INDIVIDUAL  CONTRIBUTION | Analyzing the data | Features Extraction | Graphical Representation |  |
| GUIDE | Dr. Vinothini C | | | |
|  |  | | | |
| PROJECT ABSTRACT: | We are interested in utilizing big data techniques on a database containing data about patients for whom we want to predict exposure to a brain stroke. We aim to determine the probability of a given stroke observations on patients from the dataset provided. The score determined by applying a scoring model to the quantitative and qualitative information on the person (socio-demographic data, purchasing behaviour, prior answers) will assist in identifying different reasons for the attack. | | | |
|  |  | | | |
| INTRODUCTION | Stroke is a medical disorder in which the blood arteries in the brain are ruptured, causing damage to the brain. When the supply of blood and other nutrients to the brain is interrupted, symptoms might develop. According to the World Health Organization (WHO), stroke is the greatest cause of death and disability globally. Early recognition of the various warning signs of a stroke can help reduce the severity of the stroke.  Stroke occurs when the blood flow to various areas of the brain is disrupted or diminished, resulting in the cells in those areas of the brain not receiving the nutrients and oxygen they require and dying. A stroke is a medical emergency that requires urgent medical attention. Early detection and appropriate management are required to prevent further damage to the affected area of the brain and other complications in other parts of the body.  The World Health Organization (WHO) estimates that 15 million people worldwide suffer from strokes each year, with one person dying every 4-5 minutes in the affected population. Stroke is the 16 leading cause of mortality in the United States according to the Centers for Disease Control and Prevention (CDC) [1]. Stroke is a noncommunicable disease that kills approximately 11% of the population. | | | |
|  |  | | | |
| DESIGN | 1. **DATA UNDERSTANDING**   1)id: unique identifier.  2)gender: "Male", "Female" or "Other".  3) age: age of the patient. 4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension. 5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease. 6) ever\_married: "No" or "Yes". 7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed". 8) Residence\_type: "Rural" or "Urban". 9) avg\_glucose\_level: average glucose level in blood. 10) bmi: body mass index. 11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*. 12) stroke: 1 if the patient had a stroke or 0 if not.   1. **Statistics on Specific Features**     **3)Graphical Representation**  outputheatmap | | | |
| PLATFORM USED  (H/W & S/W TOOLS TO BE USED | Jupyter Notebook, Pyspark, Seaborn | | | |
|  |  | | | |
| PROJECT SOURCE CODE LINK (GITHUB/ GOOGLE DRIVE) |  | | | |
|  |  | | | |
| CONCLUSION /FUTURE ENHANCEMENT |  | | | |
|  |  | | | |
| UI SCREENSHOTS |  | | | |