Devon Midkiff  
Dr. Weihua Zhou  
Big Data Analytics SAT5165  
10/02/2024

Cardiovascular Disease Prediction Using Spark

The goal of this project is to predict the likelihood of an individual developing cardiovascular disease based on various health-related factors. Cardiovascular diseases are a leading cause of death globally, and early prediction can significantly enhance prevention efforts. By leveraging Spark's capabilities for large-scale data processing, this project aims to preprocess and analyze health data to identify key predictors of cardiovascular disease.

The dataset chosen for this project is the Cardiovascular Disease Dataset from Kaggle. It contains over 100,000 records, with features that include demographic information, health checkup history, physical activity, and various health conditions such as heart disease, diabetes, and cancer. The dataset is uploaded to GitHub and I will keep the project updated there.

The project will include both data preprocessing and statistical analysis using Spark and Python:

Data Preprocessing:

* Handling missing values through imputation or deletion.
* Scaling numerical features like height, weight, and BMI.
* Encoding categorical variables such as smoking and alcohol consumption history.
* Removing outliers to improve model performance.

Statistical Analysis:

* Logistic Regression will be applied to predict the likelihood of cardiovascular disease. This binary classification is well-suited to the problem at hand.
* Correlation Analysis will be used to identify relationships between different health-related features and the risk of developing heart disease.

The project will be implemented using Spark and Python, running on at least two virtual machines (VMs) to simulate a distributed computing environment. The Spark MLlib library will be used for both data preprocessing and logistic regression modeling. Python libraries such as pandas and scikit-learn will also be employed.

Key Python packages:

* PySpark for distributed data processing.
* pandas for data manipulation.
* scikit-learn for logistic regression modeling.

The Spark cluster will distribute the computational workload across the VMs, ensuring efficient processing of the dataset. The final deliverables will include correlation insights and a logistic regression model for cardiovascular disease prediction.

Data Source Link:

[Cardiovascular Disease Dataset](https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset?resource=download)

[GitHub Link](https://github.com/DevonMidkiff/Cardiovascular-Disease-Prediction-Using-Spark)