**Assignment 12.2**

**1. Use the below given data set**

**Data Set**

**2. Perform the below given activities:**

**a. Apply PCA to the dataset and show proportion of variance**

**b. Perform PCA using SVD approach**

**c. Show the graphs of PCA components**

|  |
| --- |
|  |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
| # Load SparkR library into your R session | |
|  | | library(SparkR) | |
|  | |  | |
|  | | # Initialize SparkSession | |
|  | | sparkR.session(appName = "SparkR-ML-randomForest-example") | |
|  | |  | |
|  | | # Random forest classification model | |
|  | |  | |
|  | | # $example on:classification$ | |
|  | | # Load training data | |
|  | | df <- read.df("libsvm\_data.txt", source = "libsvm") | |
|  | | training <- df | |
|  | | test <- df | |
|  | |  | |
|  | | # Fit a random forest classification model with spark.randomForest | |
|  | | model <- spark.randomForest(training, label ~ features, "classification", numTrees = 10) | |
|  | |  | |
|  | | # Model summary | |
|  | | summary(model) | |
|  | |  | |
|  | | # Prediction | |
|  | | predictions <- predict(model, test) | |
|  | | head(predictions) | |
|  | |  | |
|  | |  | |
|  | | # Random forest regression model | |
|  | |  | |
|  | | # Load training data | |
|  | | df <- read.df("linear\_regression\_data.txt", source = "libsvm") | |
|  | | training <- df | |
|  | | test <- df | |
|  | |  | |
|  | | # Fit a random forest regression model with spark.randomForest | |
|  | | model <- spark.randomForest(training, label ~ features, "regression", numTrees = 10) | |
|  | |  | |
|  | | # Model summary | |
|  | | summary(model) | |
|  | |  | |
|  | | # Prediction | |
|  | | predictions <- predict(model, test) | |
|  | | head(predictions) | |
|  | |  | |
|  | |  | |
|  | | sparkR.session.stop() | |