**Big Data Programming**

**CSEE5590-0004/490-0004**

**Project Part 2 Report**

Team : 6

Chandrasekhar Pentakota (class ID : 9)

Mustafa Bilal (class ID : 15)

**Links :**

<https://github.com/chanduhub/BigDataProgramming/wiki>

<https://youtu.be/kVZaC1hl5NU>

https://github.com/bmian93/BigDataProgramming/tree/master/BigDataProgramming/ProjectPart2

**PREDICTING LOAN CREDIT RISK**

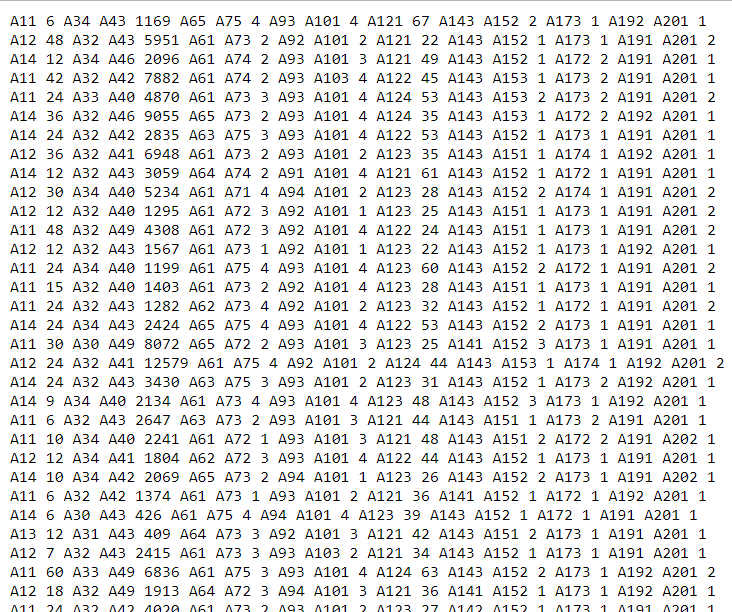
**Introduction**: Machine Learning is been part of everyday life of human beings without knowing them, suppose if we look at YouTube videos, they analyse our data and sends us the videos which we cannot skip watching them. Not only YouTube it’s been used many top companies like Google, Microsoft, Apple and many others. In our project we are going to predict bank loan credit risk using Apache sparks ml Random Forests and Decision trees.

**Background:**

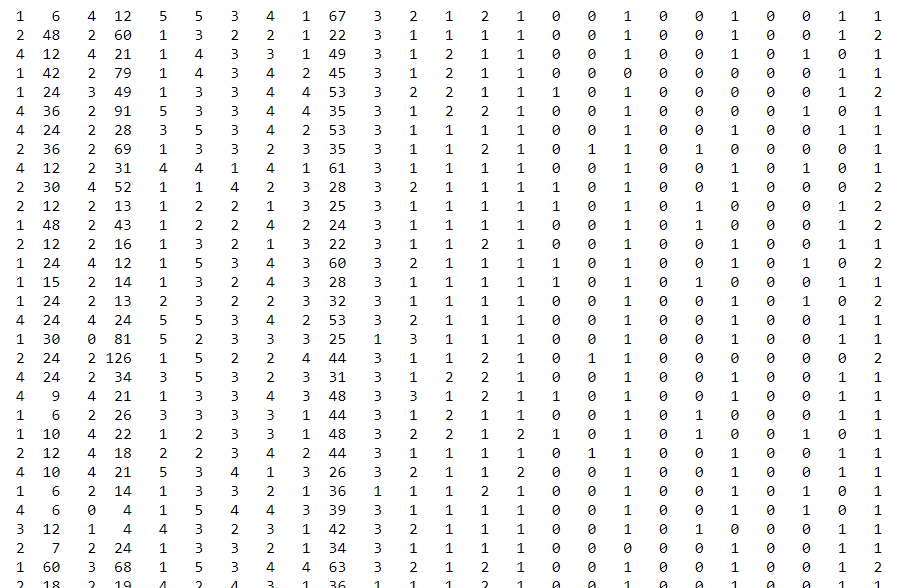
**Tools and Technologies -** Spark ML (Classification Algorithms Random Forests and Decision Tree), IntelliJ

Input data contains string columns which must be converted to integers.

Before conversion -



After conversion –



**Model:**

**Decision Tree –** A decision tree is a supervised machine learning classification algorithm that uses a tree like model of decisions and their possible outcomes. Below we can ask the first question like if checking account balance is greater than 200 and then two more questions can be derived like duration of credit greater then 29 months for partial creditability and length of employment greater than 9 months for partial non-creditability.

Checking Account Balance > 200

yes no

Length of employment > 9 months

Duration of Credit < 29 months

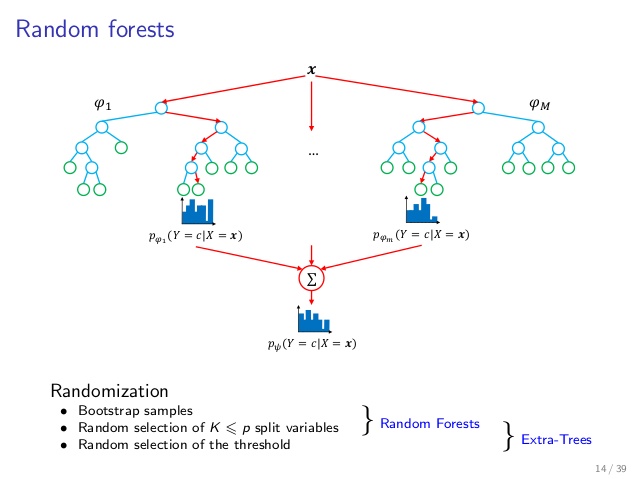
yes yes no

Payment status of previously loan = paid

yes no

**Random Forest:**

Random forests are the best classification algorithm out of all classification algorithms and it gives the more accuracy compared to other algorithms. This algorithm is a popular ensemble learning method for both classification and regression. This model consists of multiple decision trees depending on different subsets of data at the training stage.

****

Source: google

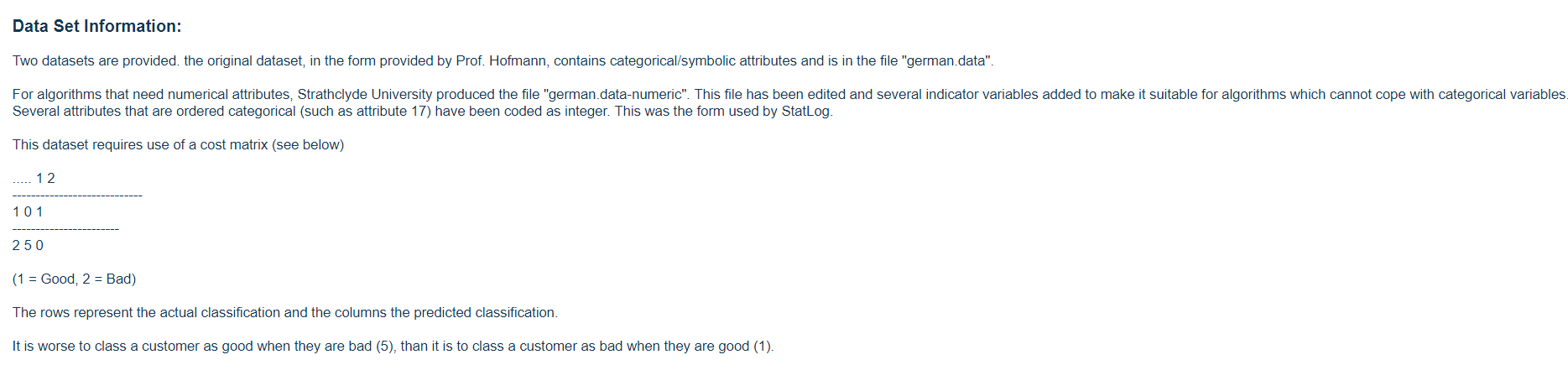
**Dataset:**

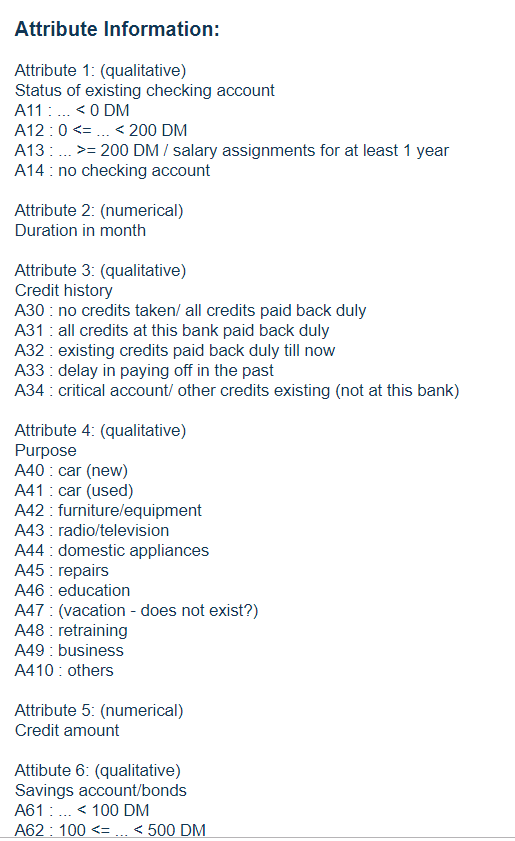
The dataset consists of 20 attributes and a total of 1000 instances or records. Attributes include both categorical and integer values.

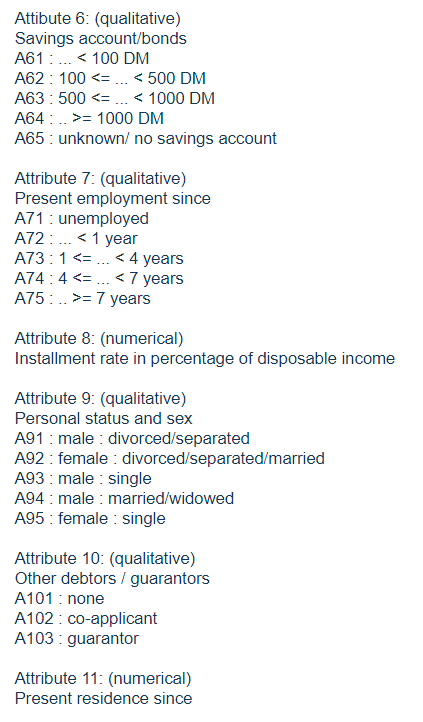
Below are the features of the data.

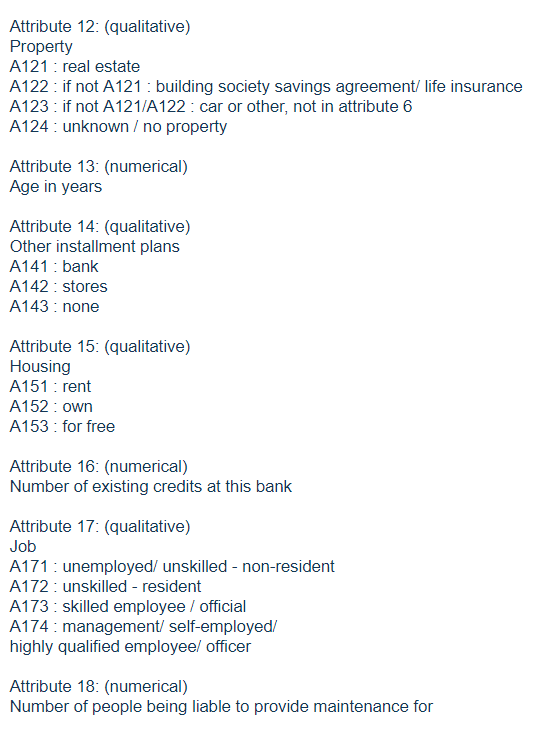
Features → {"balance", "duration", "history", "purpose", "amount", "savings", "employment", "instPercent", "sexMarried", "guarantors", "residenceDuration", "assets", "age", "concCredit", "apartment", "credits", "occupation", "dependents", "hasPhone", "foreign"}

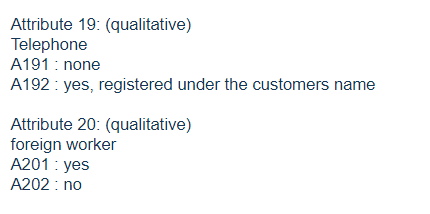
Dataset and attribute information –











**Analysis of Data:**

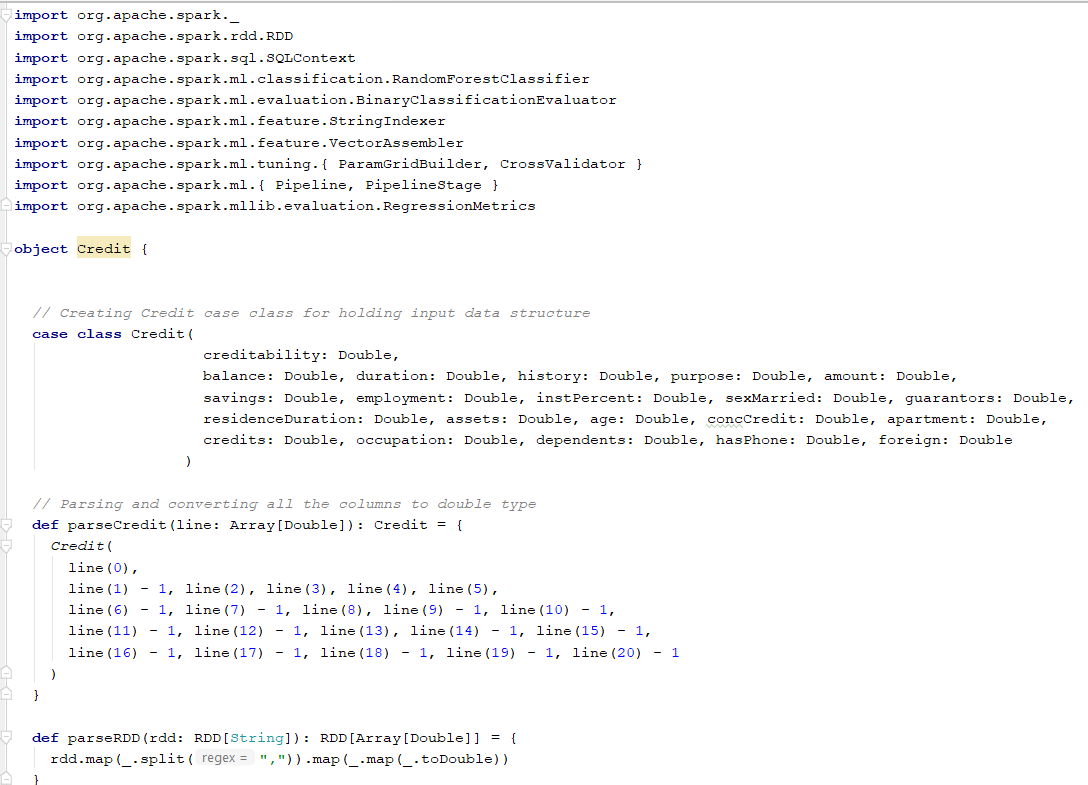
Input data from the csv file is read as RDD and it is parsed and then converted to data frame which makes life easier to analyse.Also, we have used SparkSQL for pranalysis of data which shows the complete description about the column like mean, max and min values of that column. Finally, we have applied Machine learning models to predict whether the bank customer is credit worthy or not to give a loan.

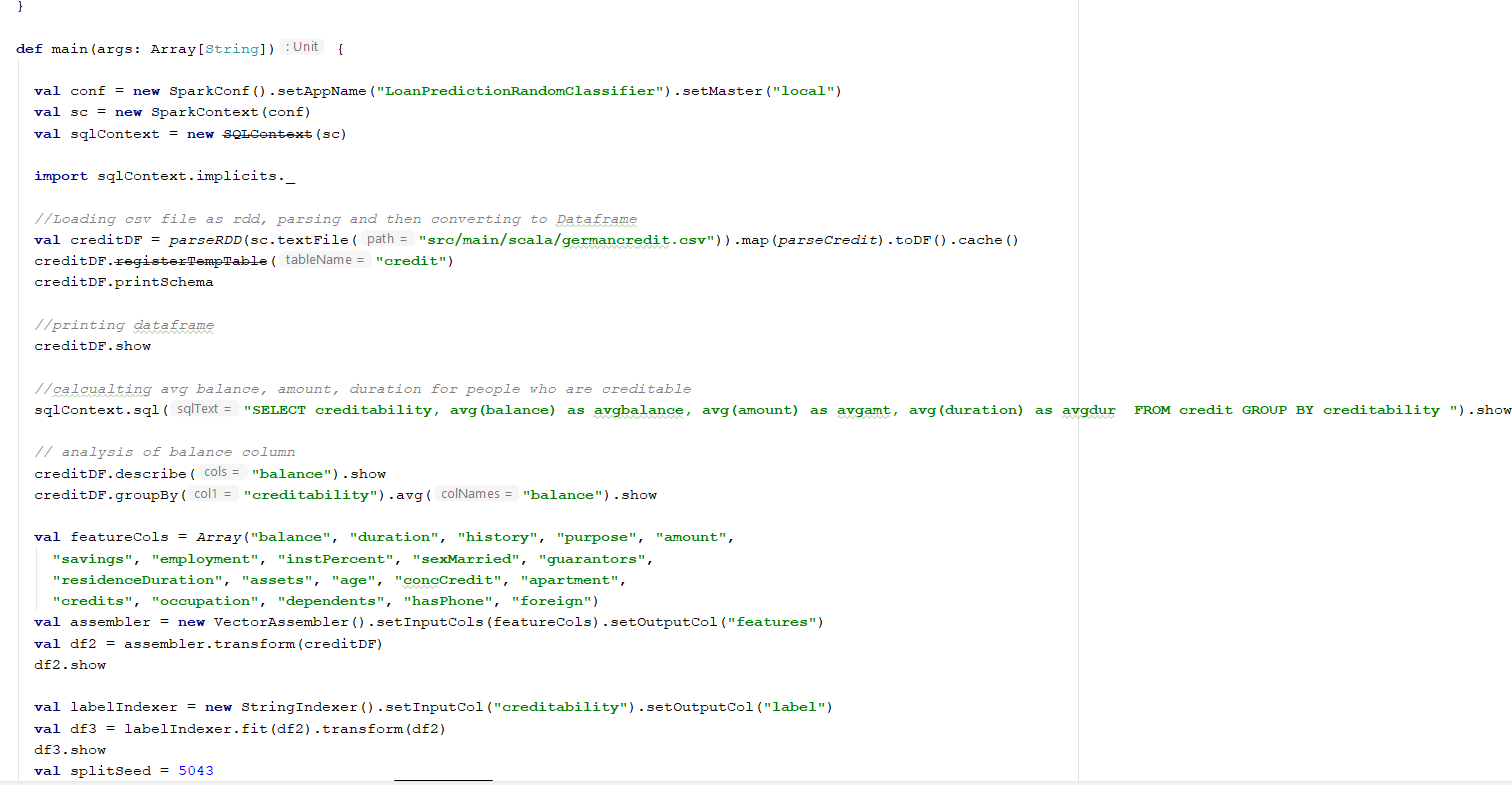
Lectures used are Spark RDD, Spark Data Frames, SparkSQL and Spark Machine Learning.

**Implementation:**

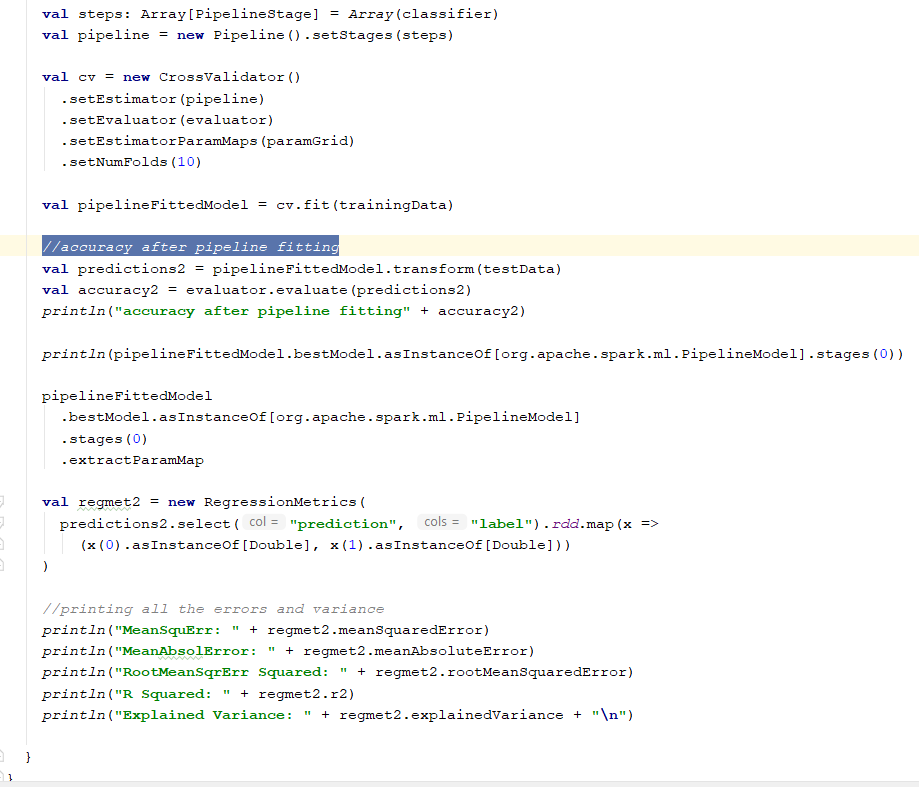
Code is implemented in Spark Machine Learning using both random forest and decision tree classifiers.

Random Forest code –









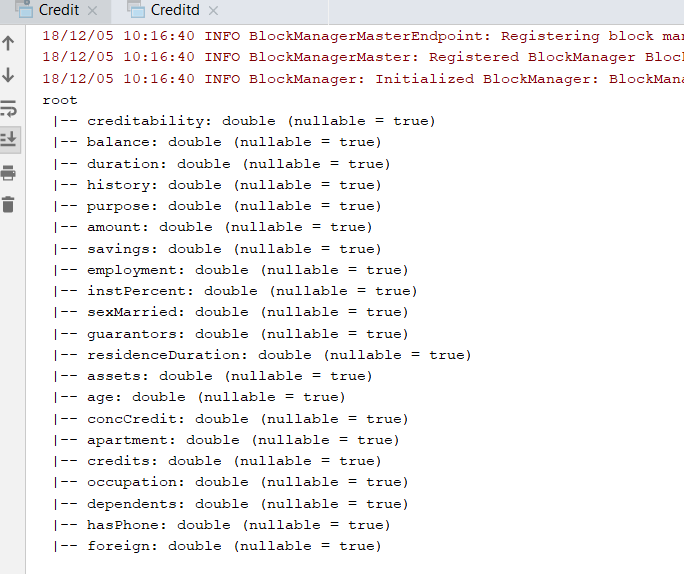
Decision Tree Code –

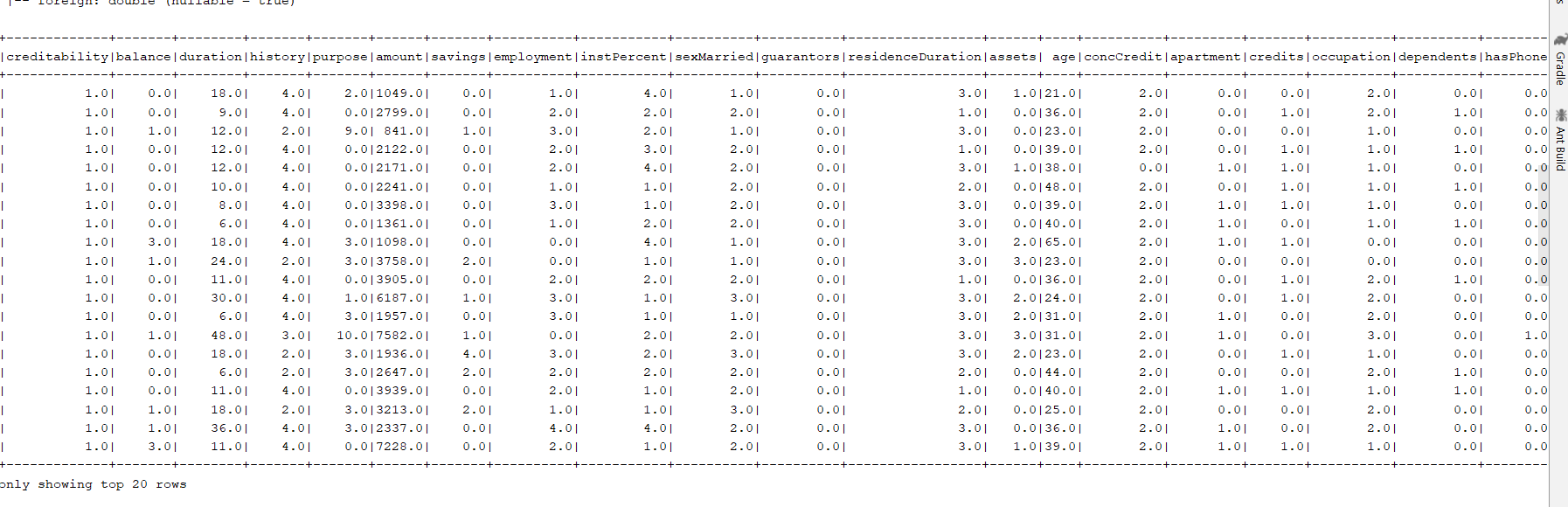


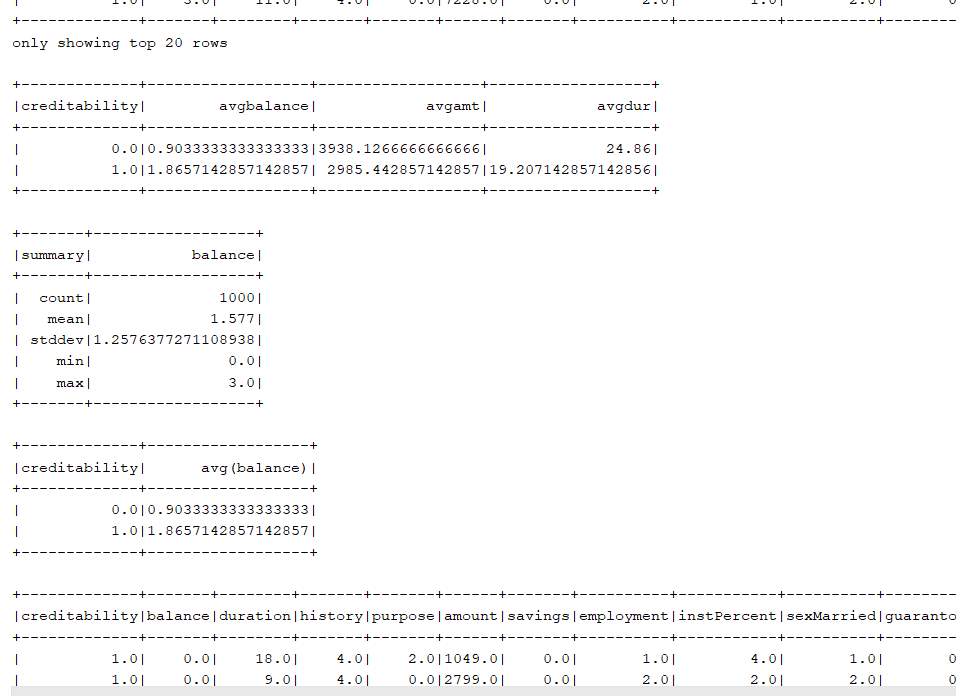
**Results Evaluation:**

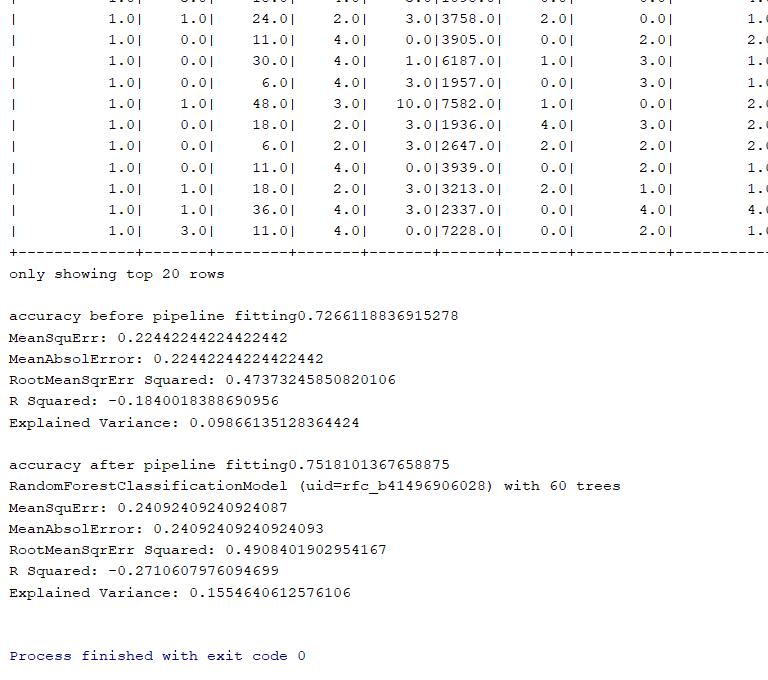
By comparing both the models, we can say that Random Forest model is the best of two, which has the highest accuracy among them and less error compared to the other model.

Random Forest output –

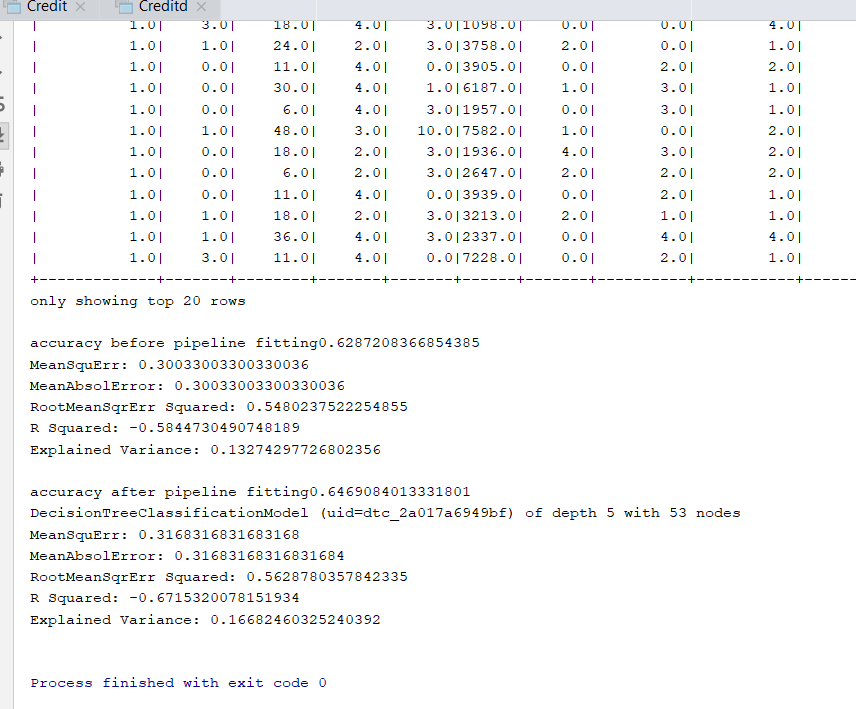








Decision Tree output –



**Conclusion:**

In conclusion Random Forest gives better results than the decision tree. Also, by using ML Pipeline gives higher accuracy percentage than without the ML Pipeline. A pipeline gives a basic method to experiment with various mixes of parameters, utilizing a procedure called network look, where you set up the parameters to test, and MLLib will test every one of the mixes. Pipelines make it simple to tune a whole model building work process without a moment's delay, instead of tuning every component in the Pipeline independently.

**Project Management:**

Implementation status report

Work completed: 100%

Description:

German bank loan Dataset has been loaded and format and feature has been identified/extraction.

Applied ML models on the identified dataset, calculated accuracy of the models and tested them.

Responsibility (Task, Person)

Downloading and analysing Dataset: Chandra sekhar Pentakota

Worked on Random Forest: Chandra sekhar Pentakota

Worked on Decision Tree: Bilal Mustafa

Contributions (members/percentage):

Chandra sekhar Pentakota 50%

Bilal Mustafa 50%

Responsibility (Task, Person)

Apache Spark Setup / Chandra sekhar Pentakota

Reading Dataset and converting to Data Frame / Bilal Mustafa

Extract Features / Bilal Mustafa

Spark Analysis generating result / Chandra sekhar Pentakota

Training and Testing Model / Chandra sekhar Pentakota

Issues/Concerns:

None

**References/Bibliography:**

https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

https://mapr.com/blog/predicting-loan-credit-risk-using-apache-spark-machine-learning-random-forests/