SparkScala Final Project -

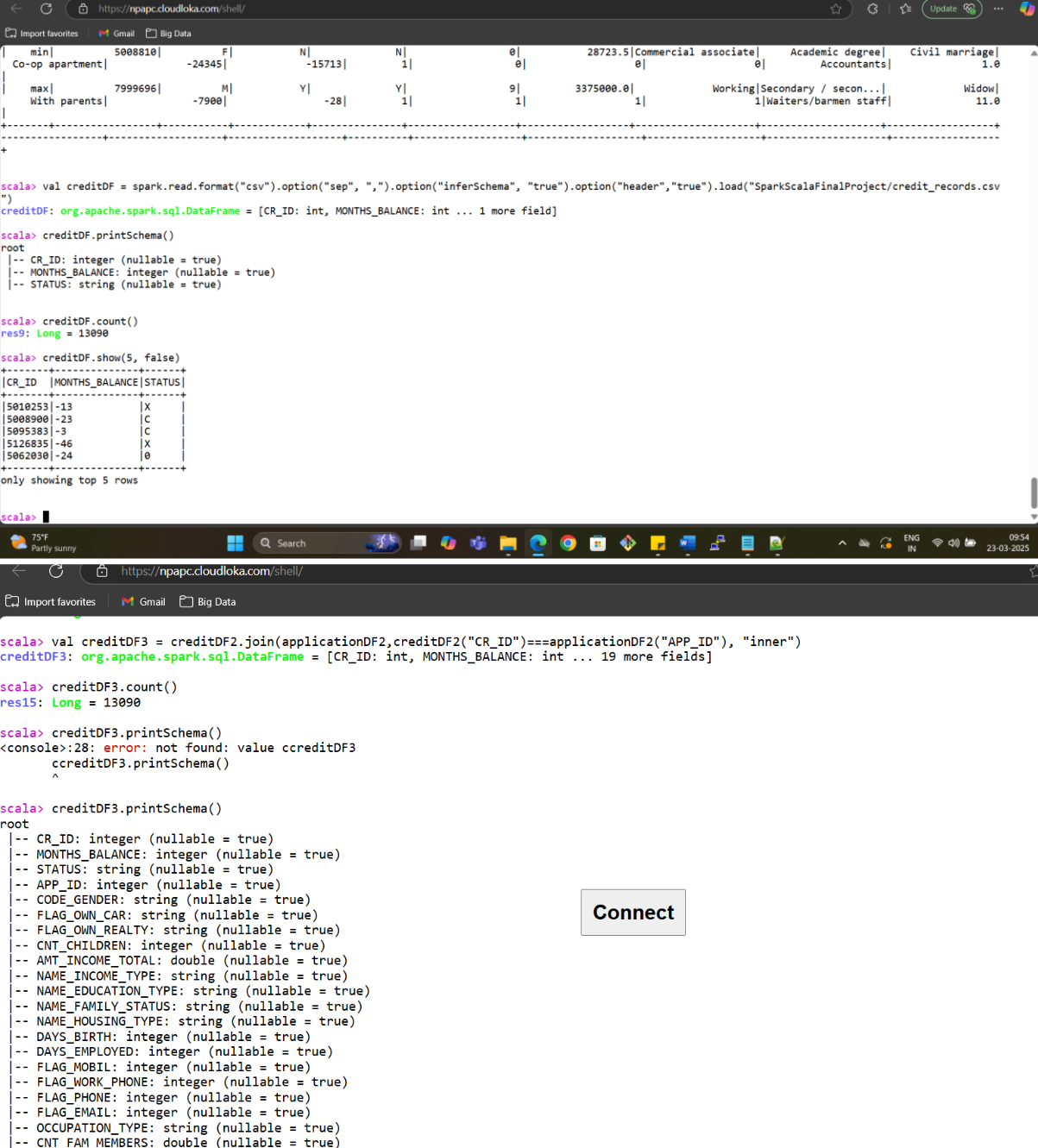
***Credit Card Approval Application in Spark***

**1. Introduction**

This document outlines the implementation of a Decision Tree Classifier in Spark ML, including key data processing steps, model training, and evaluation of its performance metrics.

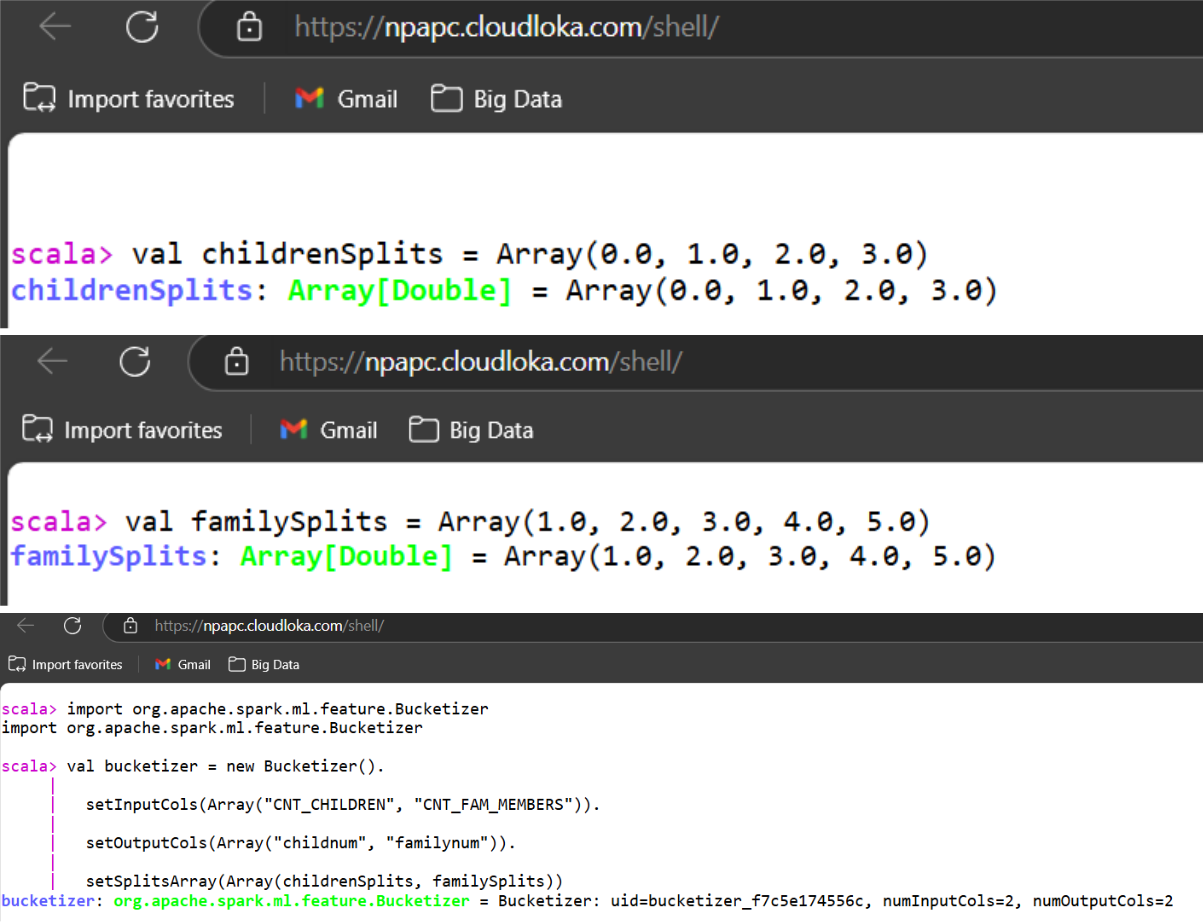
1. **Data Preparation**
   * Loading the dataset into respective data frames and performing the initial checks.





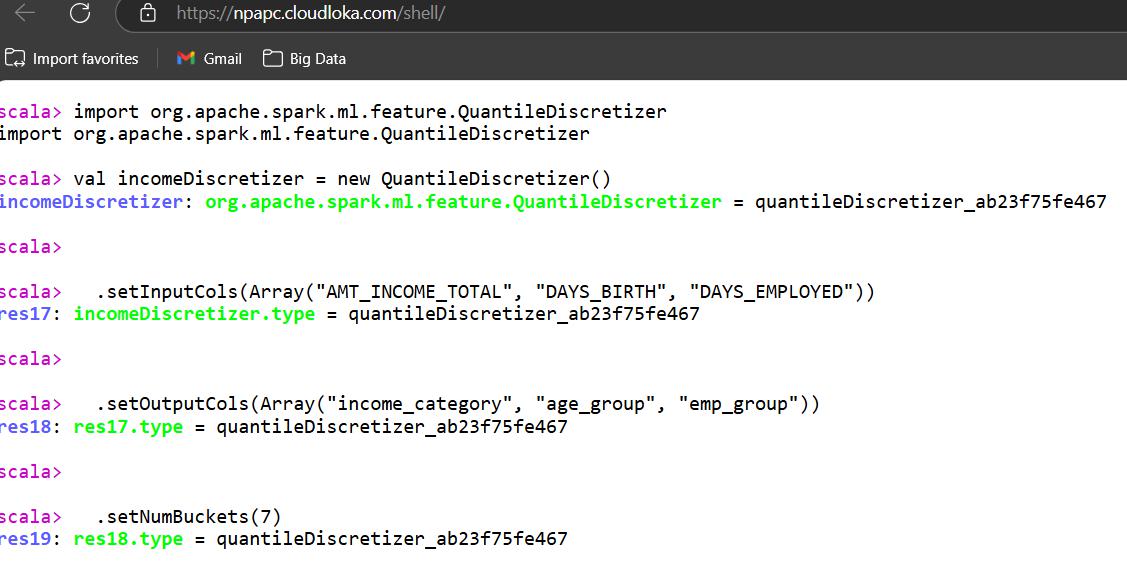
3. **Data Transformation**: Applying following transformations to the data

**Bucketizer:** Binned numerical columns into categorical values based on predefined splits.

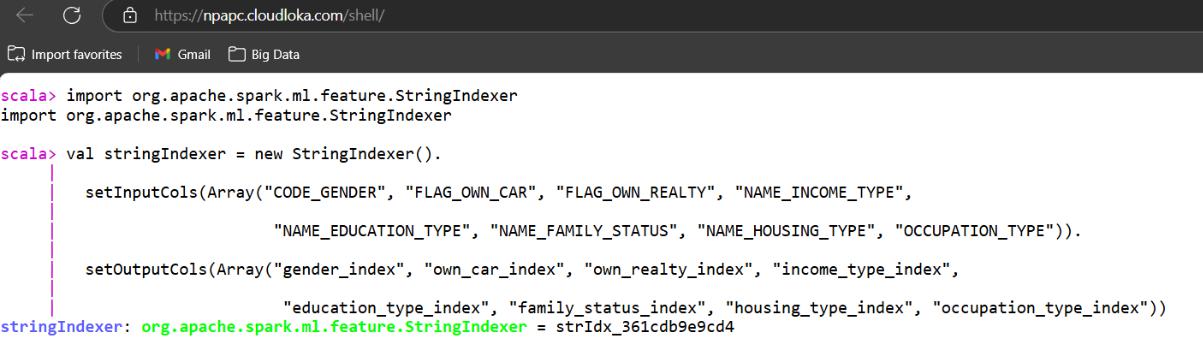


**QuantileDiscretizer**: Divided continuous numeric columns into equal-sized quantile-based bins for:

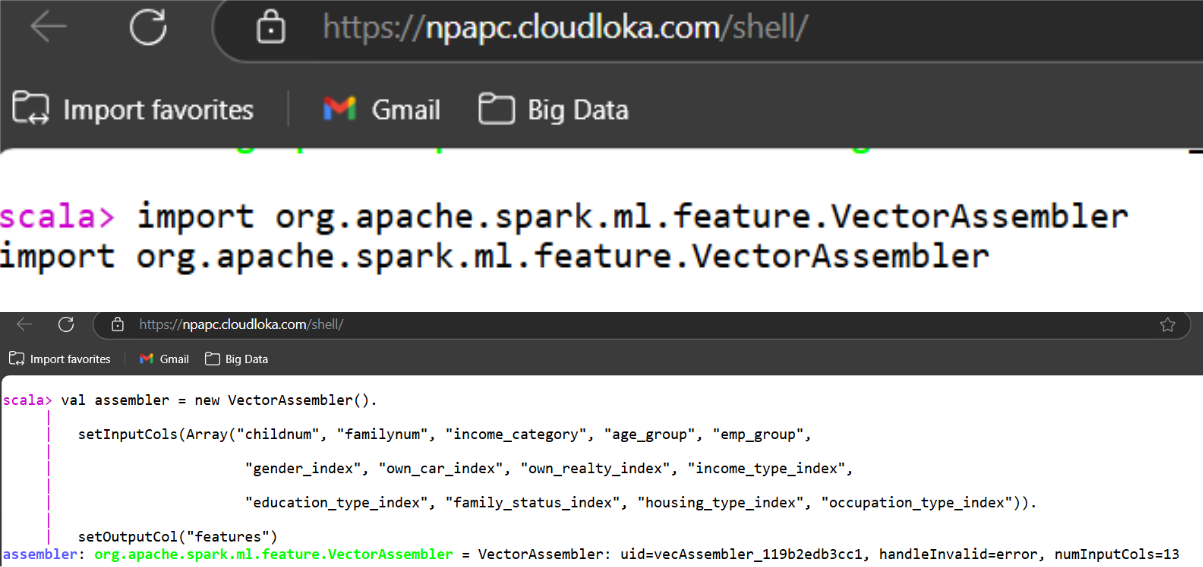
* Income: Created income categories.
* Age: Transformed days of birth into age groups.
* Employment Duration: Derived employment duration groups.



**StringIndexer:** Encoded categorical columns into numerical indices.



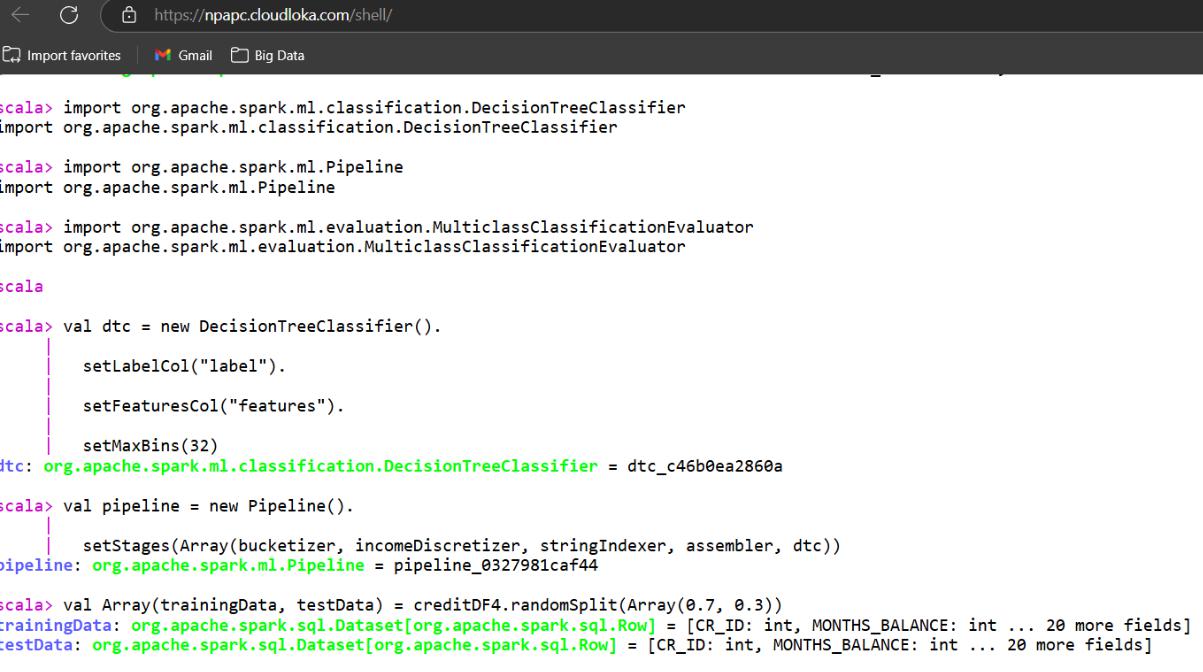
**Vector Assembler**: Combined multiple feature columns into a single vector column for model input.



**4. Model Building and Training**

**Algorithm**: Decision Tree Classifier from Spark ML.

Split the dataset into training and test sets and trained the model using the training data.



**5. Model Evaluation:**

**Metric Used**:

* Evaluated model accuracy using MulticlassClassificationEvaluator.
* Metric name: accuracy.



**6. Results**

Accuracy = 0.9946126218573628 (Perfect predictions made by the classifier).

Test Error = 0.005387378142637234

**7. Conclusion**

The Decision Tree Classifier demonstrated exceptional performance on the Credit Card Approval dataset, achieving an **accuracy of 99.46%** and a **test error of only 0.54%**. These results indicate that the model was able to classify applicants with near-perfect precision, making it highly reliable for predicting 'good' or 'bad' clients.