**LAPORAN**

**RENCANA TUGAS MANDIRI (RTM) Ke-5**

**MATA KULIAH BIG DATA C**

**“AUTOMATIC ESSAY SCORING WITH PYSPARK”**



**DISUSUN OLEH:**

Mohamad Ibnu Fajar Maulana (21083010106)

**DOSEN PENGAMPU:**

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

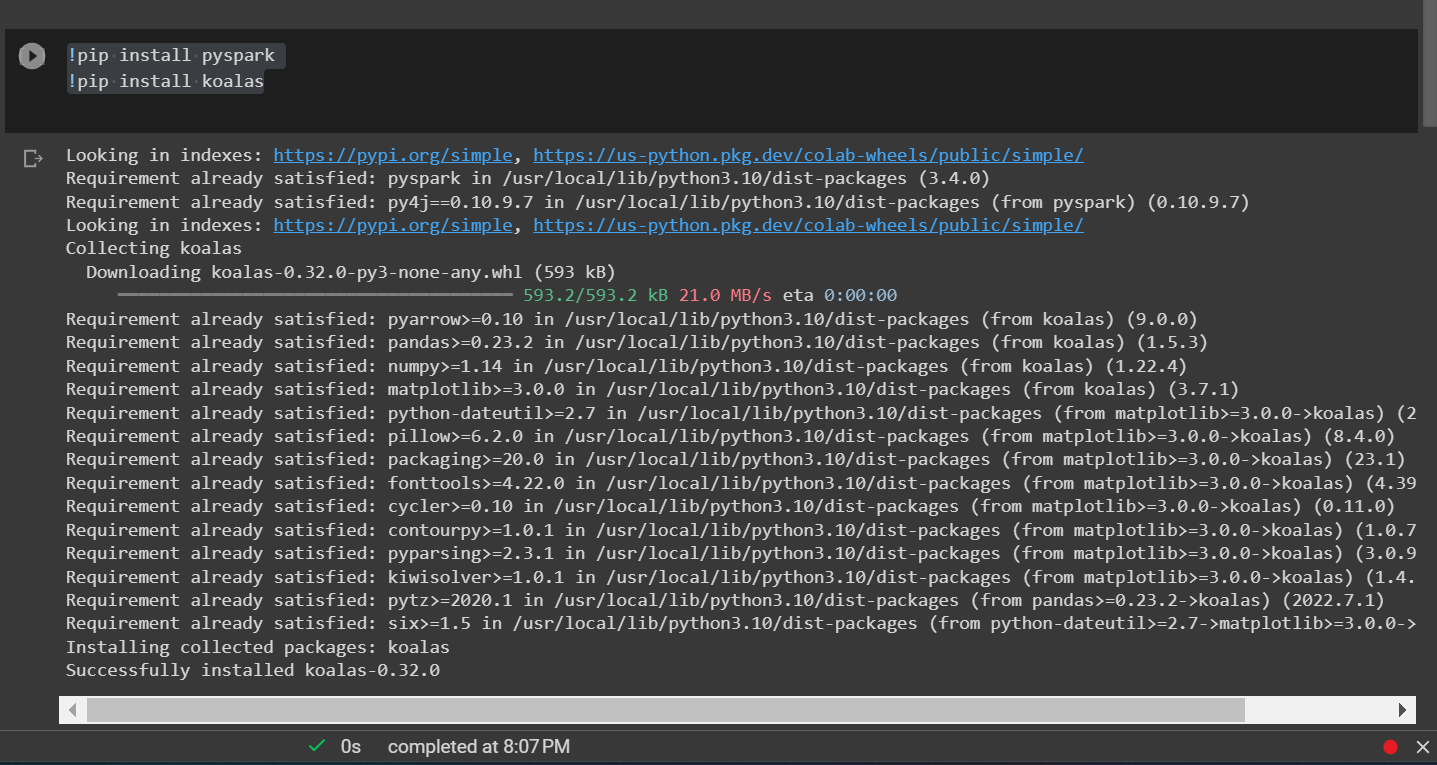
**PROGRAM STUDI SAINS DATA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR**

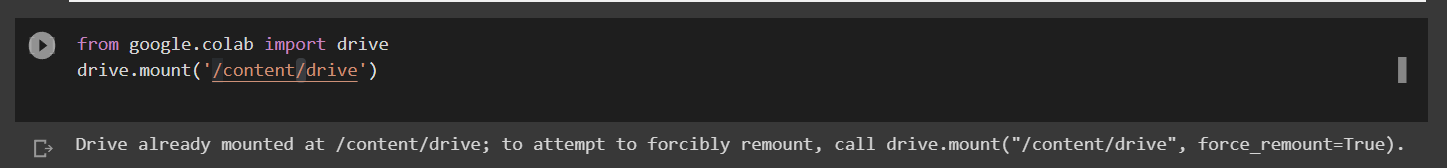
**2022**

1. **Instalasi Pyspark**

****

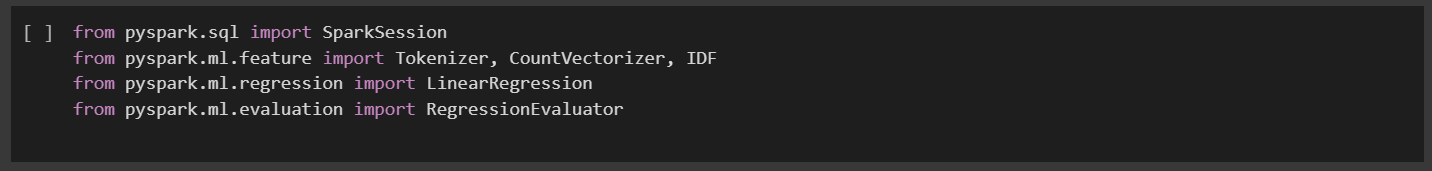
* **pyspark** adalah pustaka yang memungkinkan pengguna untuk mengakses Apache Spark menggunakan bahasa pemrograman Python. Apache Spark sendiri adalah kerangka kerja yang dirancang untuk memproses data secara terdistribusi, dengan dukungan untuk berbagai bahasa pemrograman dan jenis data.
* **koalas** adalah pustaka yang memungkinkan pengguna untuk menggunakan antarmuka Pandas untuk bekerja dengan data terdistribusi di Apache Spark. Dengan Koalas, pengguna dapat menulis kode Pandas biasa dan menggunakannya untuk memproses data yang jauh lebih besar daripada yang dapat ditangani oleh satu mesin.

1. **Pre-Processing**

****

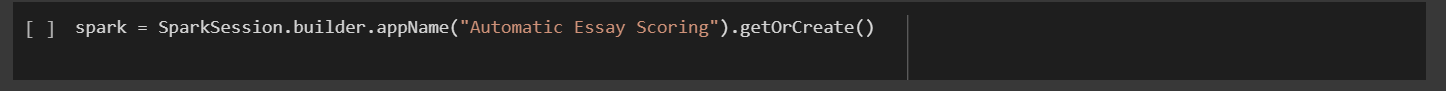
google.colab adalah modul Python yang menyediakan alat-alat dan utilitas untuk menjalankan kode di Colab. Dalam kode di atas, kita mengimpor fungsi drive dari modul google.colab, yang digunakan untuk mengakses Google Drive dari Colab.

Selanjutnya, kita memanggil fungsi mount dan memberikan argumen /content/drive kepadanya. Fungsi ini akan meminta izin akses ke Google Drive dan akan menghasilkan tautan yang harus diikuti pengguna untuk memberikan izin tersebut. Setelah izin diberikan, Colab akan dapat mengakses Google Drive melalui direktori /content/drive.

****

* pyspark.sql adalah modul yang menyediakan API untuk memproses data dalam format tabel menggunakan Spark SQL.
* SparkSession adalah kelas yang menyediakan antarmuka untuk mengakses Spark dalam pengolahan data dengan PySpark.
* pyspark.ml.feature adalah modul yang menyediakan API untuk fitur pemrosesan dan transformasi data.
* Tokenizer adalah salah satu fitur Transformer yang digunakan untuk memecah teks menjadi token.
* CountVectorizer adalah fitur Transformer yang digunakan untuk menghitung frekuensi kemunculan kata dalam dokumen.
* IDF adalah Transformer yang digunakan untuk menghitung bobot kata dalam dokumen berdasarkan frekuensi kemunculannya.
* pyspark.ml.regression adalah modul yang menyediakan API untuk model regresi.
* LinearRegression adalah salah satu algoritma regresi yang umum digunakan.
* pyspark.ml.evaluation adalah modul yang menyediakan API untuk evaluasi model.
* RegressionEvaluator adalah salah satu evaluator yang digunakan untuk menghitung metrik evaluasi dalam model regresi.

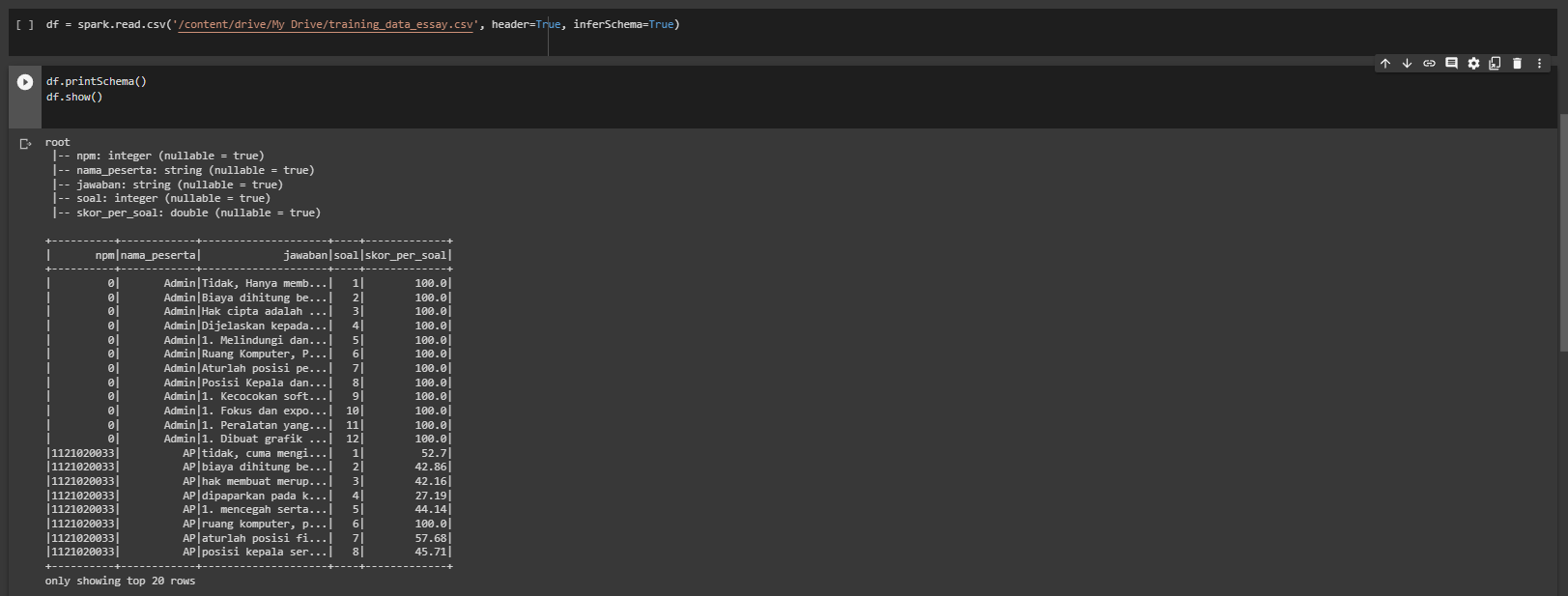
Dengan mengimpor modul ini, pengguna dapat melakukan pemrosesan data, transformasi fitur, dan membangun model regresi dalam Apache Spark menggunakan bahasa pemrograman Python.

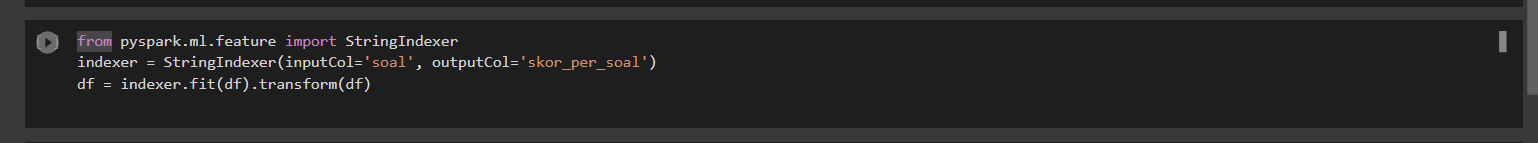


* SparkSession adalah kelas yang menyediakan antarmuka untuk mengakses Spark dalam pengolahan data dengan PySpark. Dalam kode di atas, kita menggunakan metode builder untuk membuat objek SparkSession.
* Pada objek SparkSession, kita memberikan nama aplikasi dengan metode appName("Automatic Essay Scoring"). Nama aplikasi ini akan muncul pada antarmuka pengguna Spark.

Jika sudah ada sesi Spark yang berjalan, maka sesi tersebut akan digunakan. Jika tidak ada, maka akan dibuat sesi baru dengan nama aplikasi yang diberikan. Hal ini dilakukan dengan memanggil metode getOrCreate(). Dalam praktiknya, objek spark ini akan digunakan untuk membaca dan memproses data dalam format tertentu, membangun model machine learning, dan melakukan evaluasi kinerja pada model.

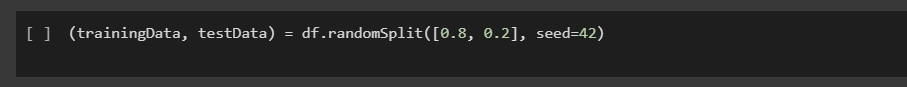
* Kemudian kita akan menampilkan DataFrame



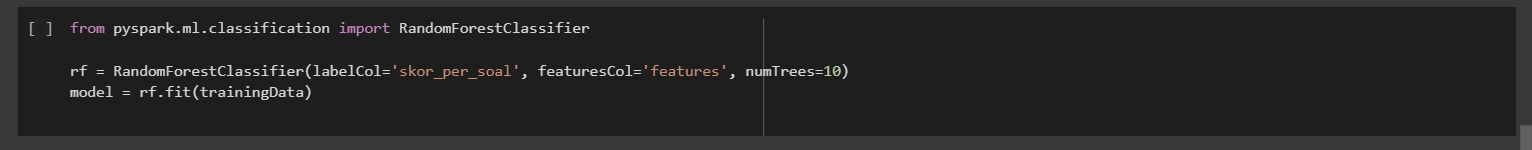
* Gambar diatas menampilkan DataFrame dengan Pyspark
* Selanjutnya 
* Kemudian melakukan training data jawaban(essay)



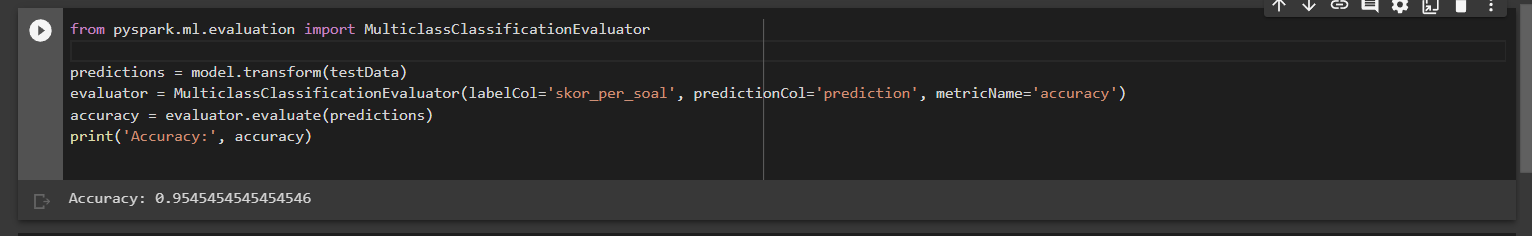
* Dimana nantinya akan dibagi data training 80% dan data testing 20%]



* Selanjutnya



* Gambar diatas melakukan automatic system scoring dengan pyspark clasification dengan merandom value
* Selanjutnya



* Menghitung ketepatan skor\_per\_soal dengan automaatic essay scoring yang mana mendapatkan Accuracy: 0.9545454545454546