**TUGAS KELOMPOK**

**PROGRAM STUDI INDEPENDEN**

**ORBIT FUTURE ACADEMY**

**Identitas Kelompok**

Kelompok : 10

Nama Anggota : Halomoan Filipus Simarmata (Jupyter XXI)

Diana Eka Riyani (Jupyter XXI)

Nyayu Chika Marselina (Jupyter XXI)

Sukma Imelda (Cordoba)

Athiya Shinta Wulandari (Cordoba)

Coach : Ipin Sugiyarto

Program : Foundations of AI and Life Skills for Gen-Z

Hari, Tanggal : Selasa, 29 Maret 2022

**Tugas:** Silahkan melakukan Data Preprocessing dengan menggunakan referensi data yg lebih banyak (Sumber bebas silahkan tentukan sendiri). sampai hasil akhir berupa akurasi menggunakan model KNN.

**Penyelesaian:**

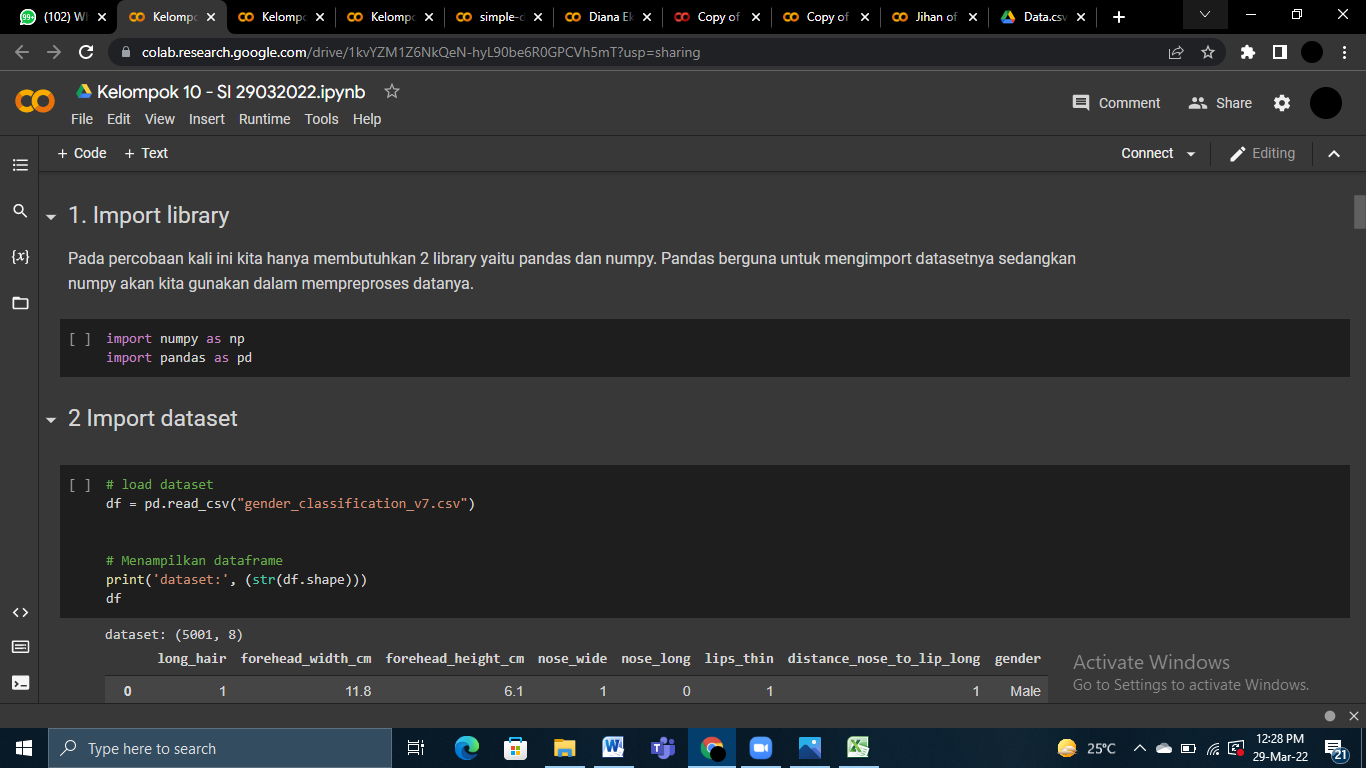
**Data Pre-Processing**

Why preprocessing?

Data yang kita punya dalam dunia nyata biasanya:

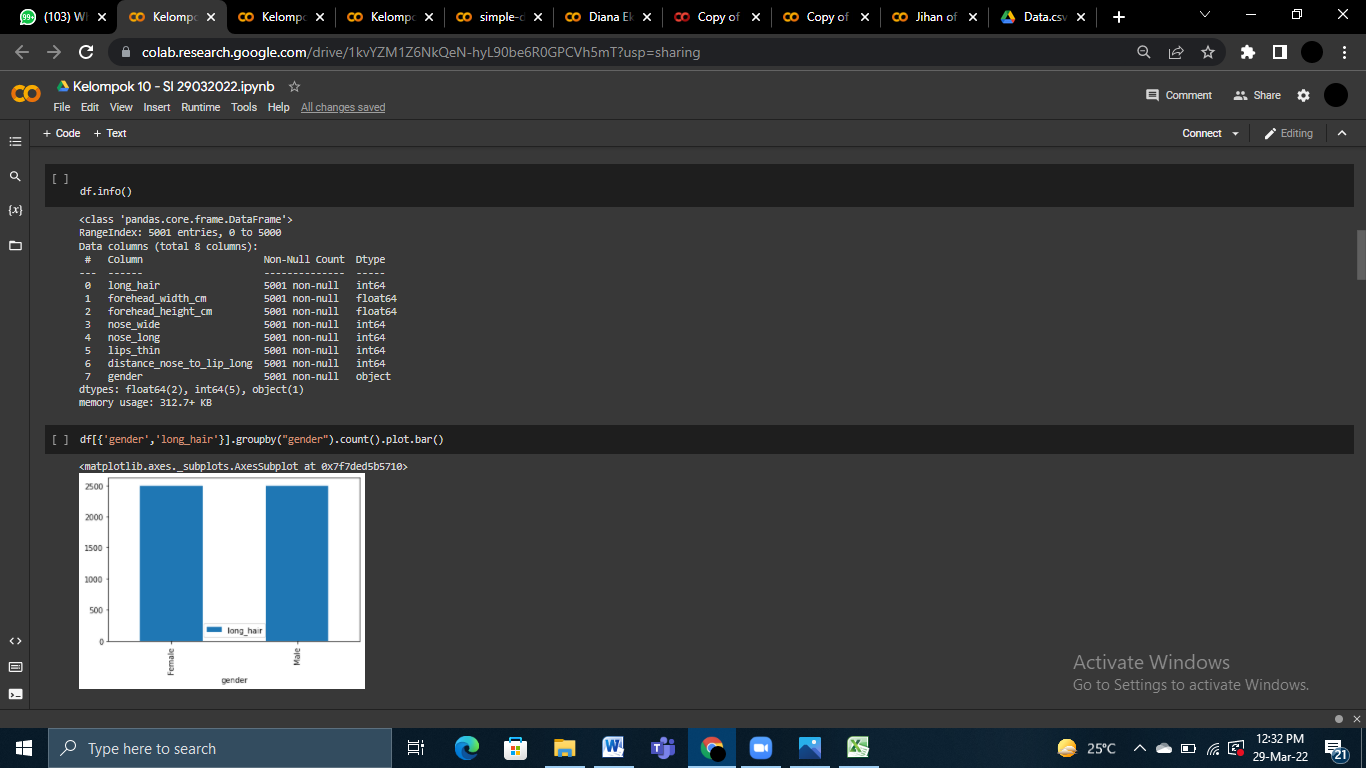
* Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
* Noisy: mengandung error seperti missing values dan outliers.
* Inconsistent: mengandung perbedaan dalam penulisan kolom

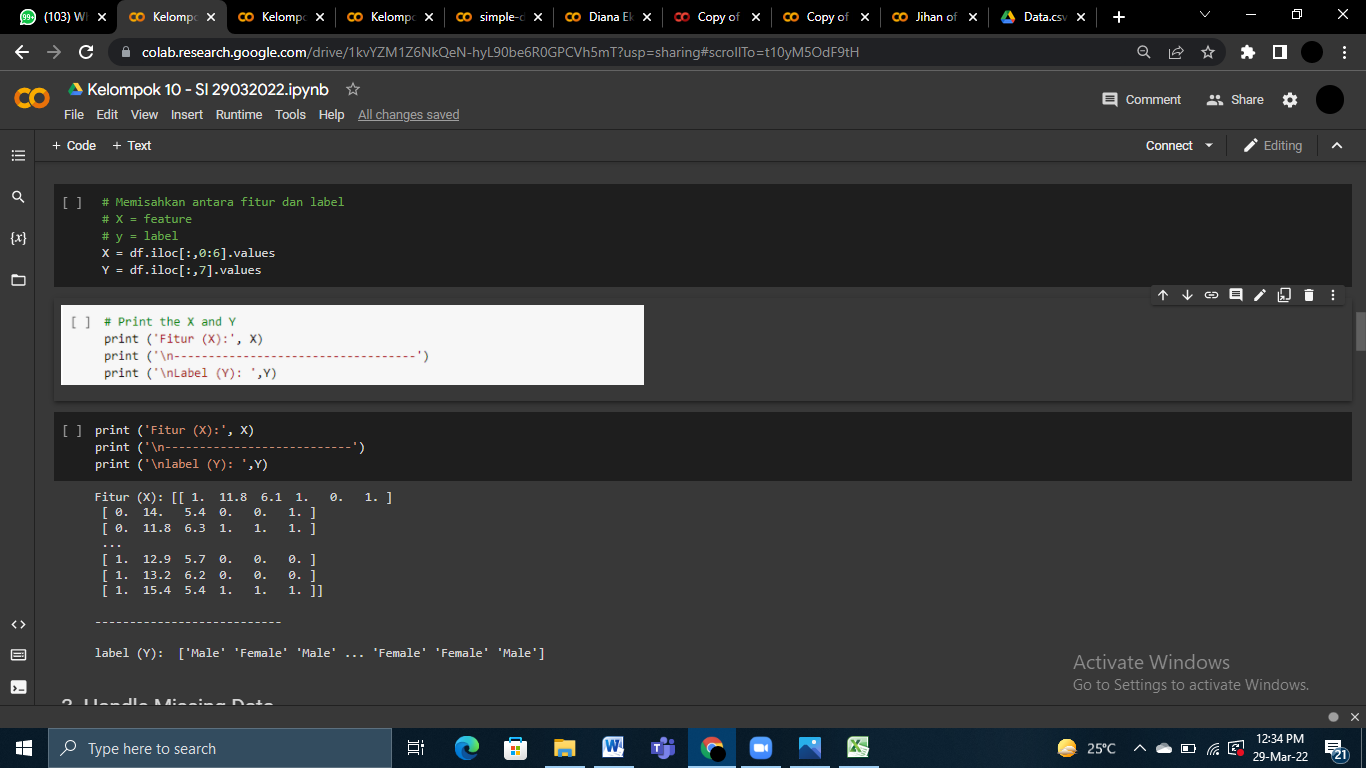
1. **Import Library**



1. **Import Dataset**

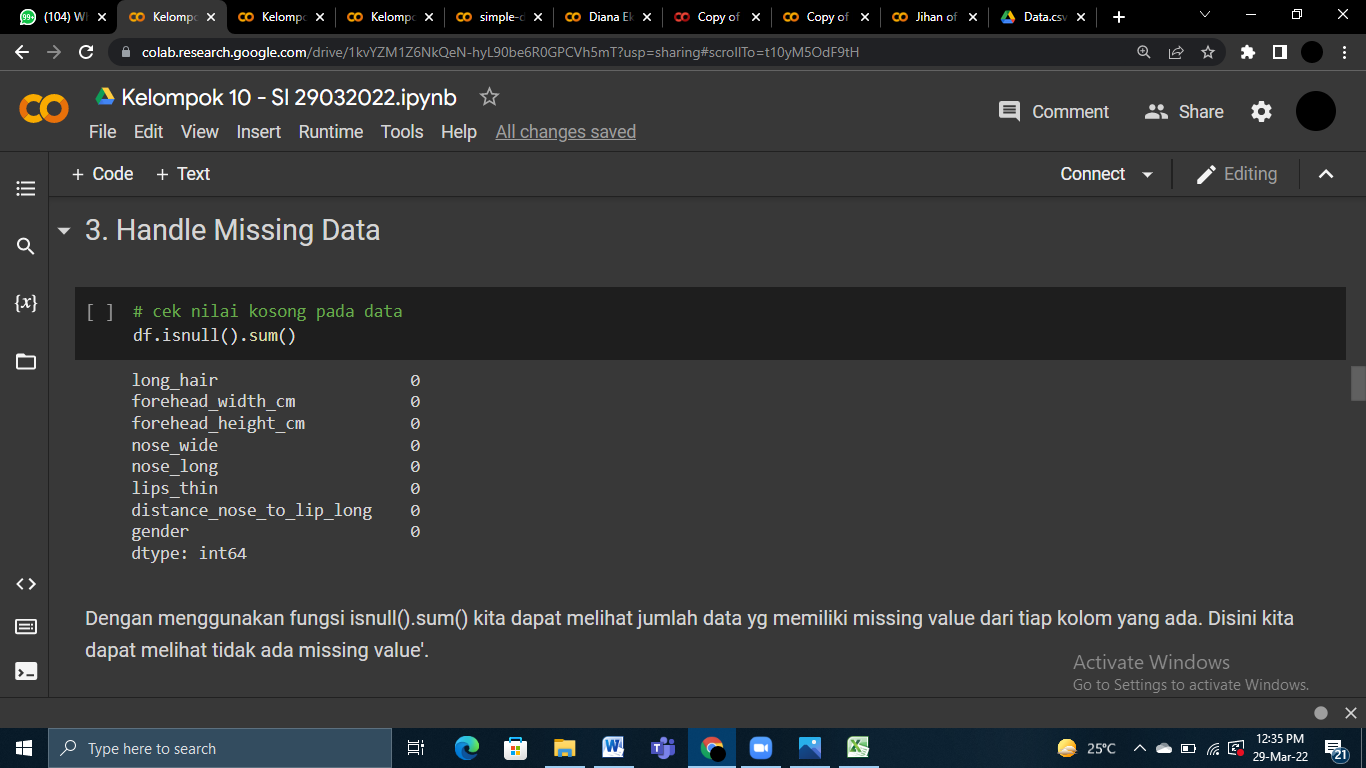
Mengimport dataset yang digunakan, di sini kelompok 10 menggunakan dataset pengklasifikasian gender. Dataset tersebut mengandung 5001 rows dan 8 columns, artinya terdapat 5001 data dan 8 kolom.

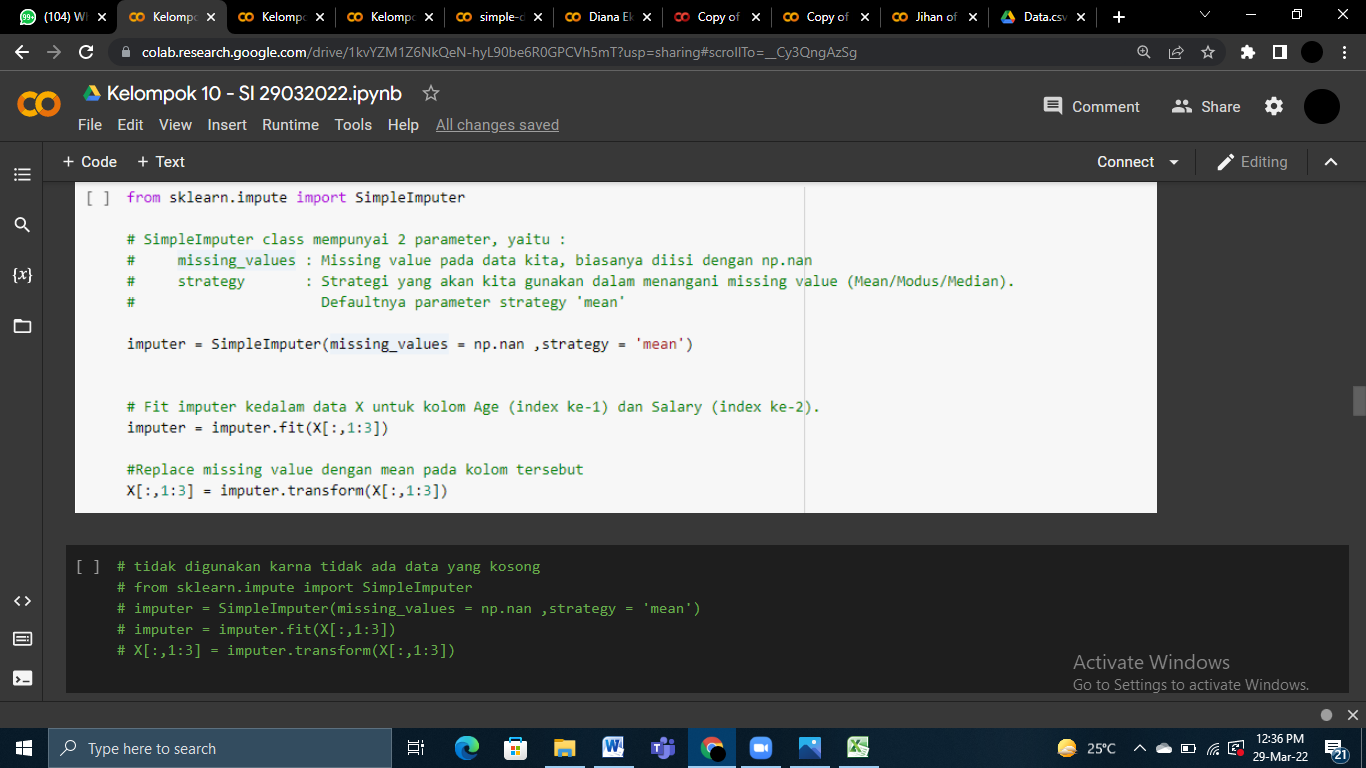


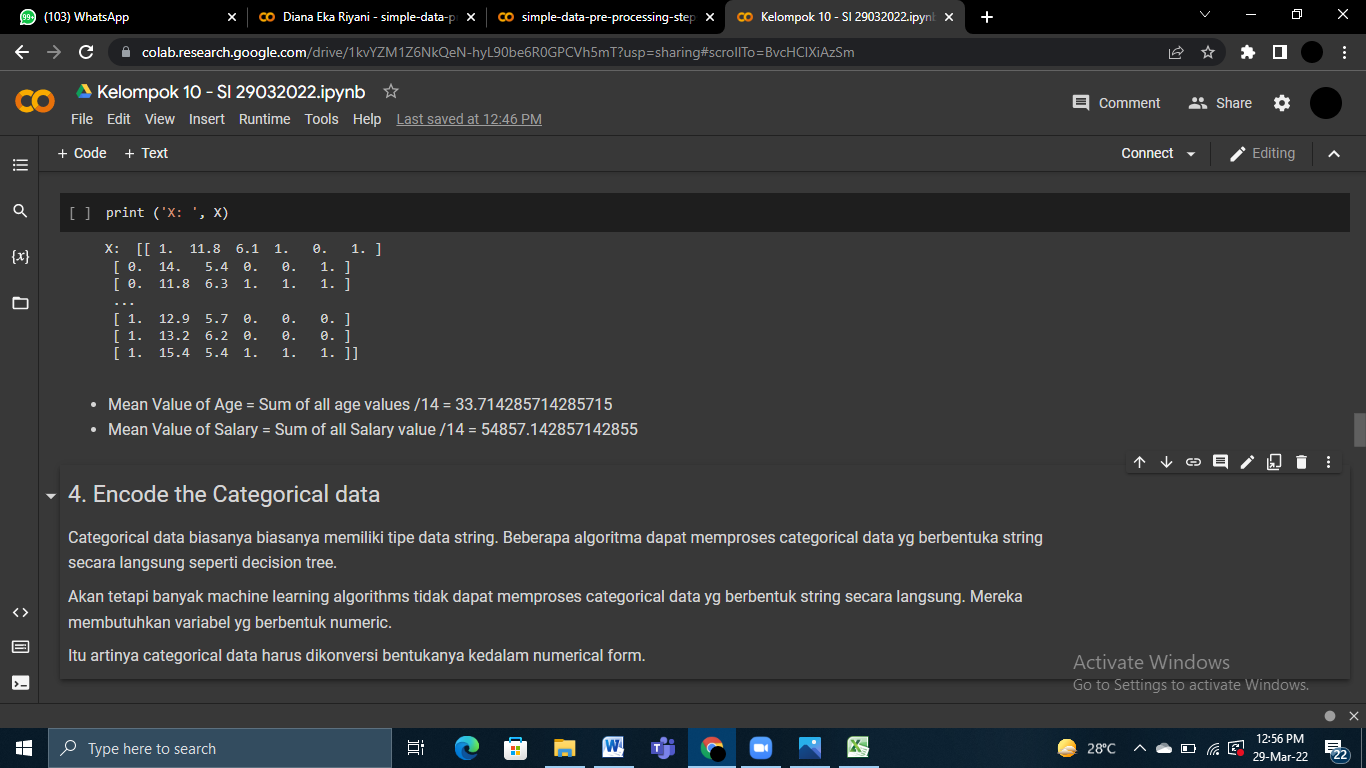


1. **Handle Missing Data**

Dengan menggunakan fungsi isnull.sum() kita dapat melihat jumlah data yang memiliki missing value dari 8 kolom. Dapat terlihat, tidak ada yang memiliki masalah missing value.

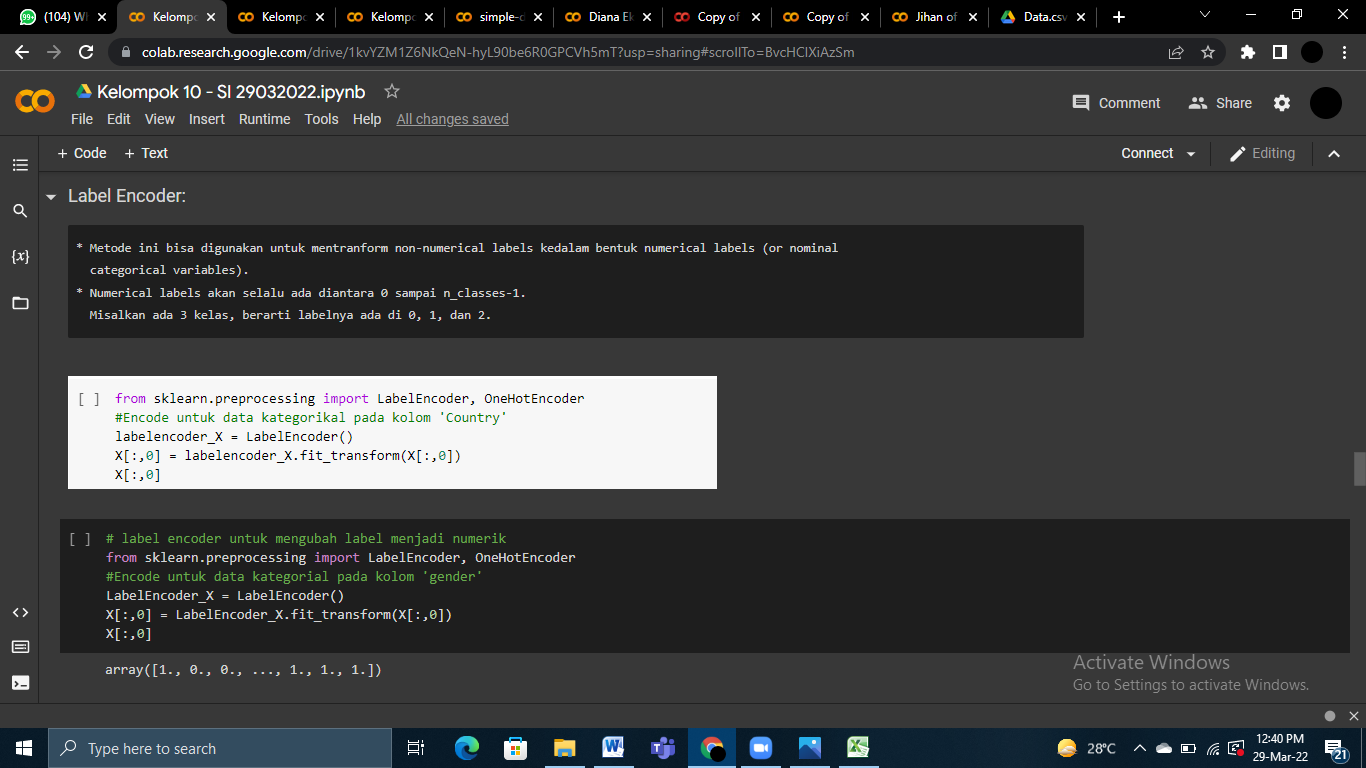


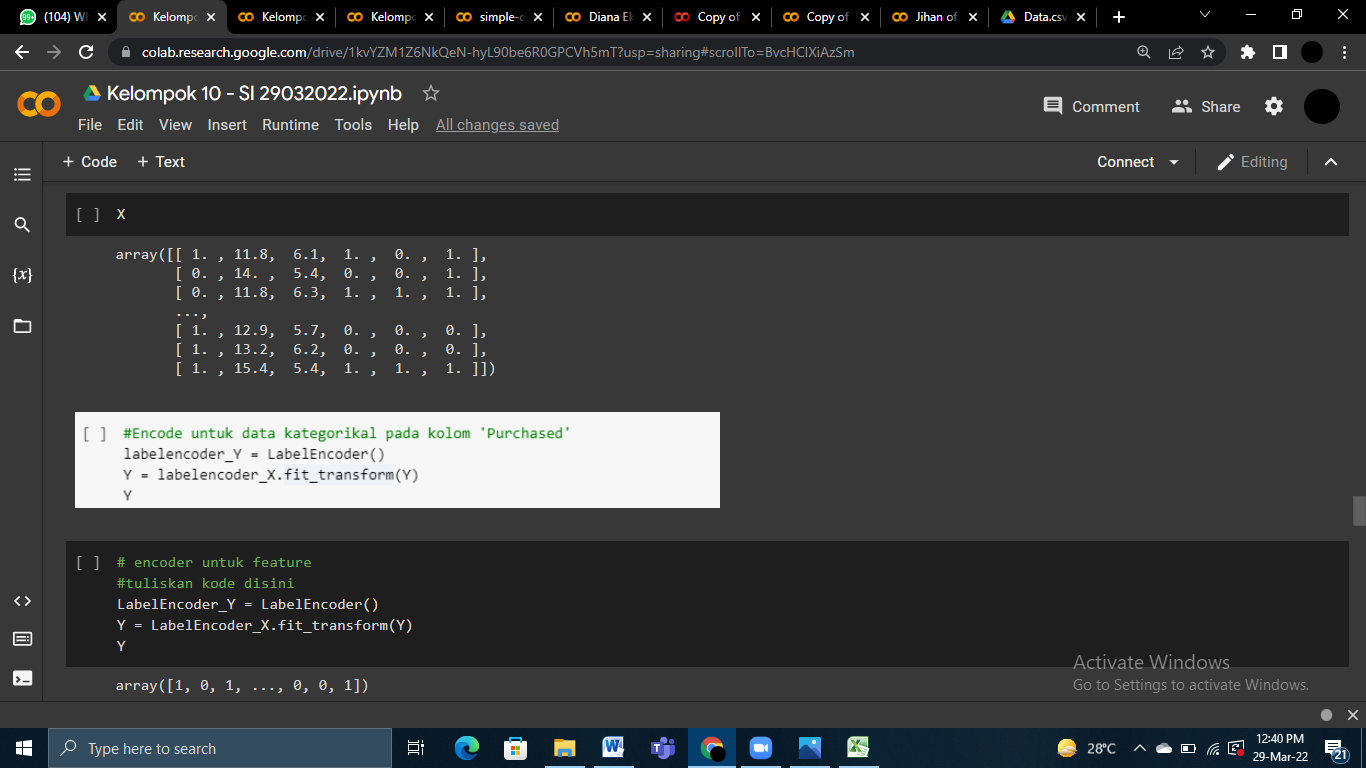




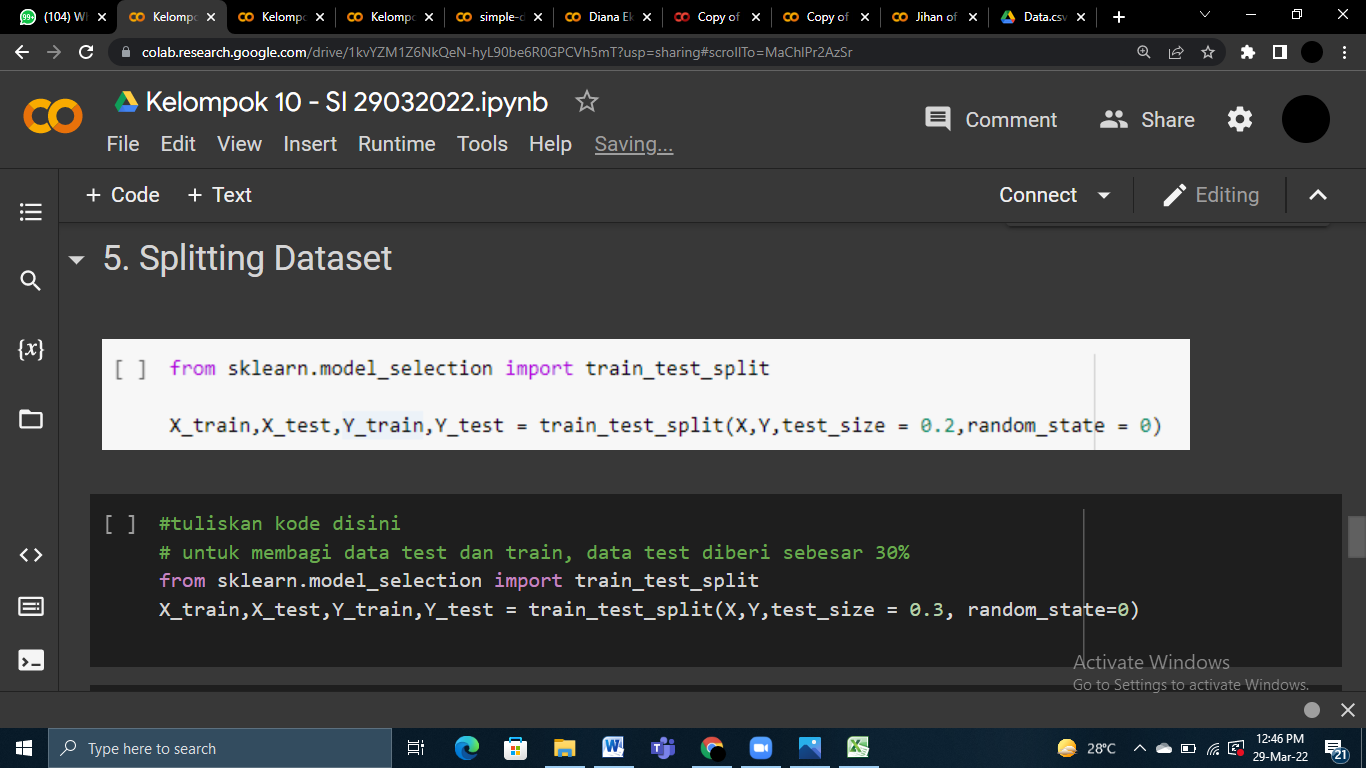
1. **Encode the Categorical data**

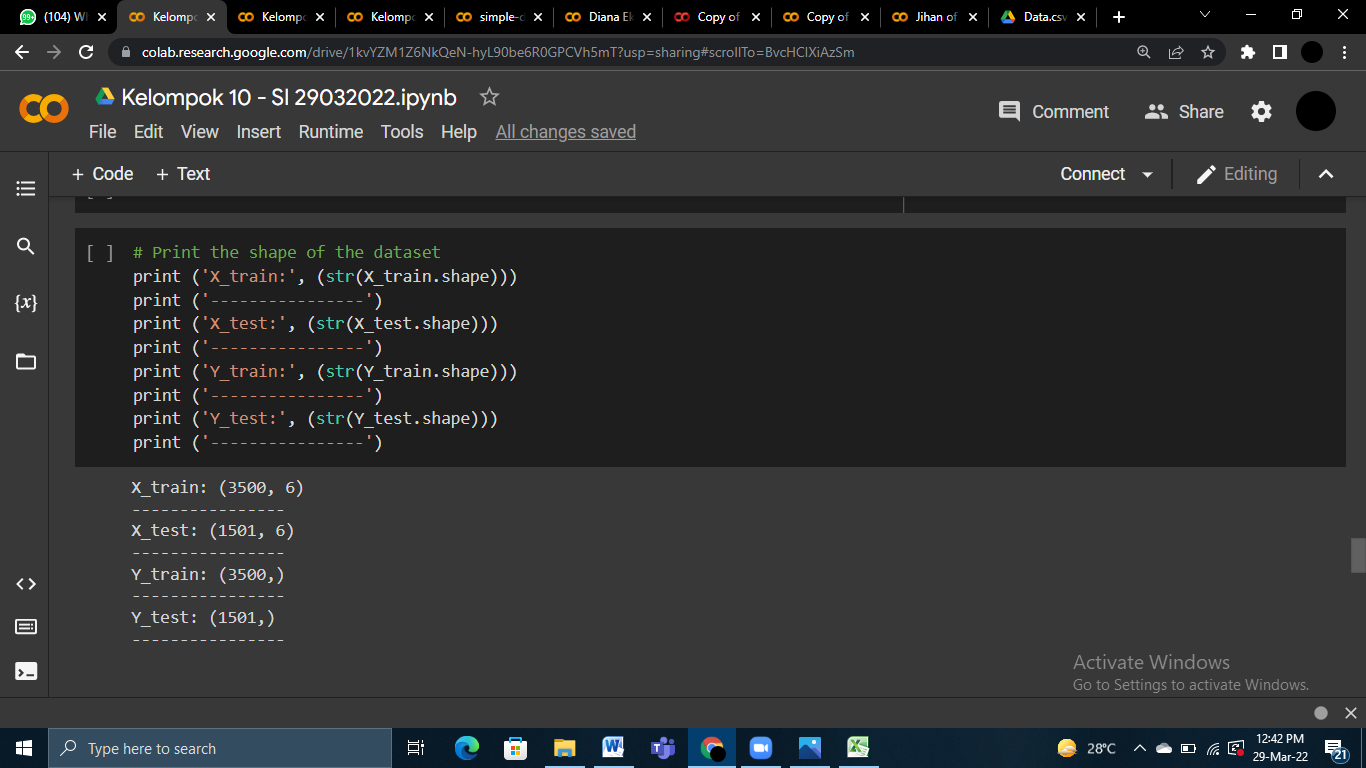
Categorical data biasanya biasanya memiliki tipe data string. Beberapa algoritma dapat memproses categorical data yang berbentuk string secara langsung seperti decision tree. Akan tetapi, banyak machine learning algorithms tidak dapat memproses categorical data yang berbentuk string secara langsung. Mereka membutuhkan variabel yang berbentuk numeric. Itu artinya categorical data harus dikonversi bentuknya ke dalam numerical form.





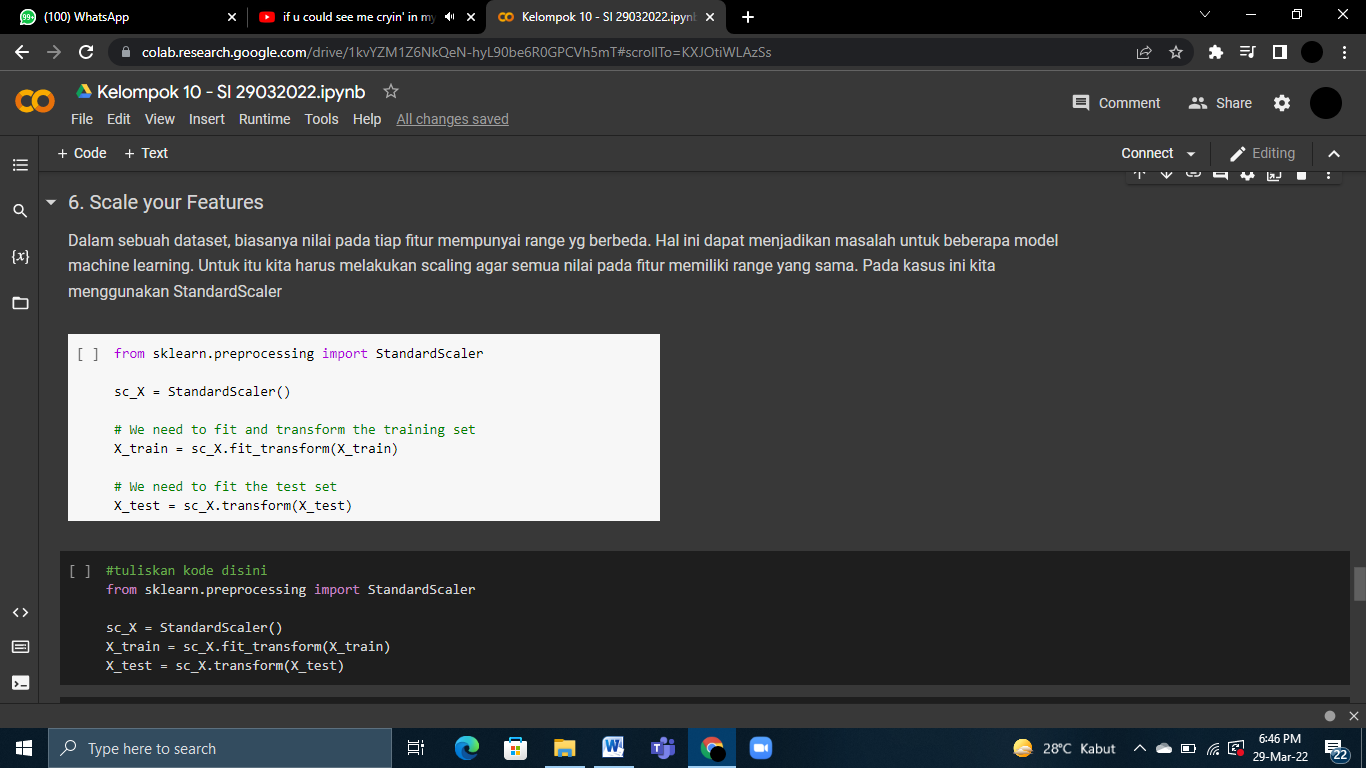
1. **Splitting Dataset**

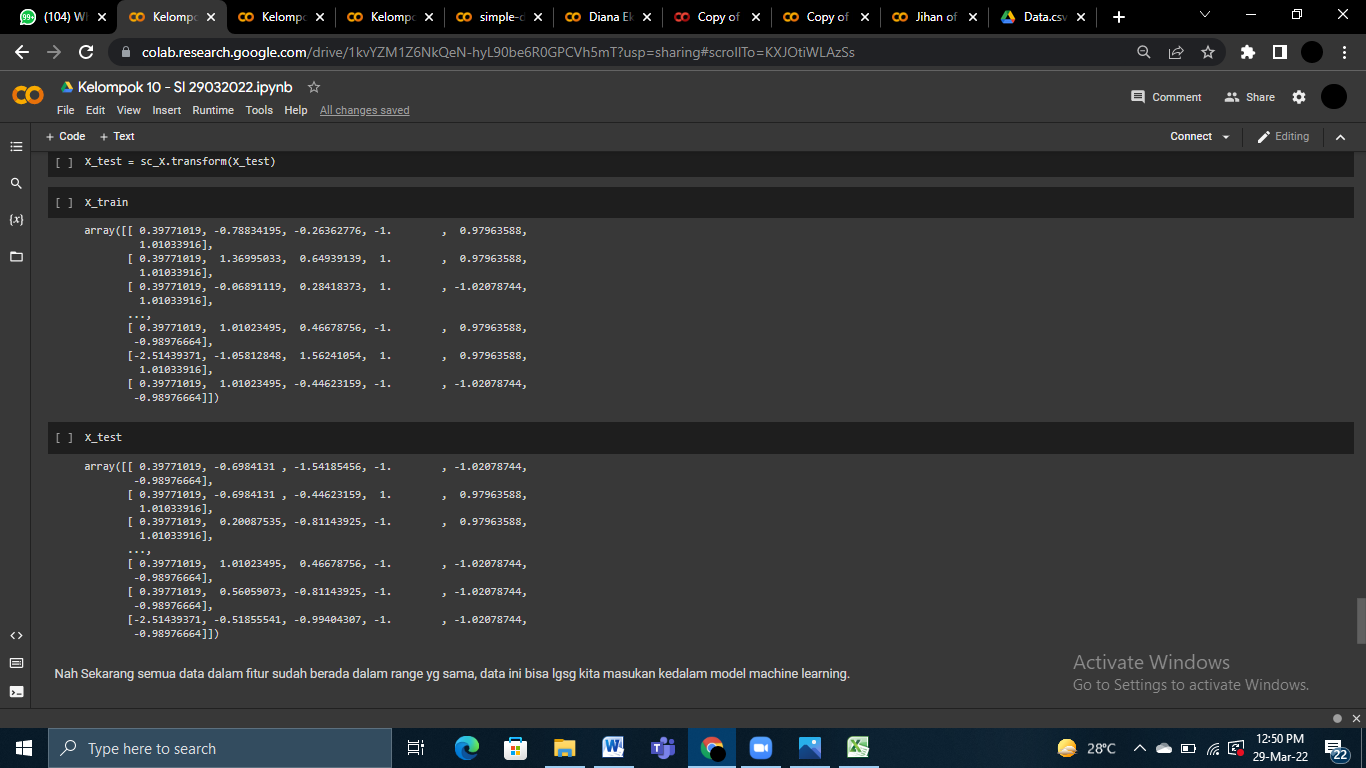


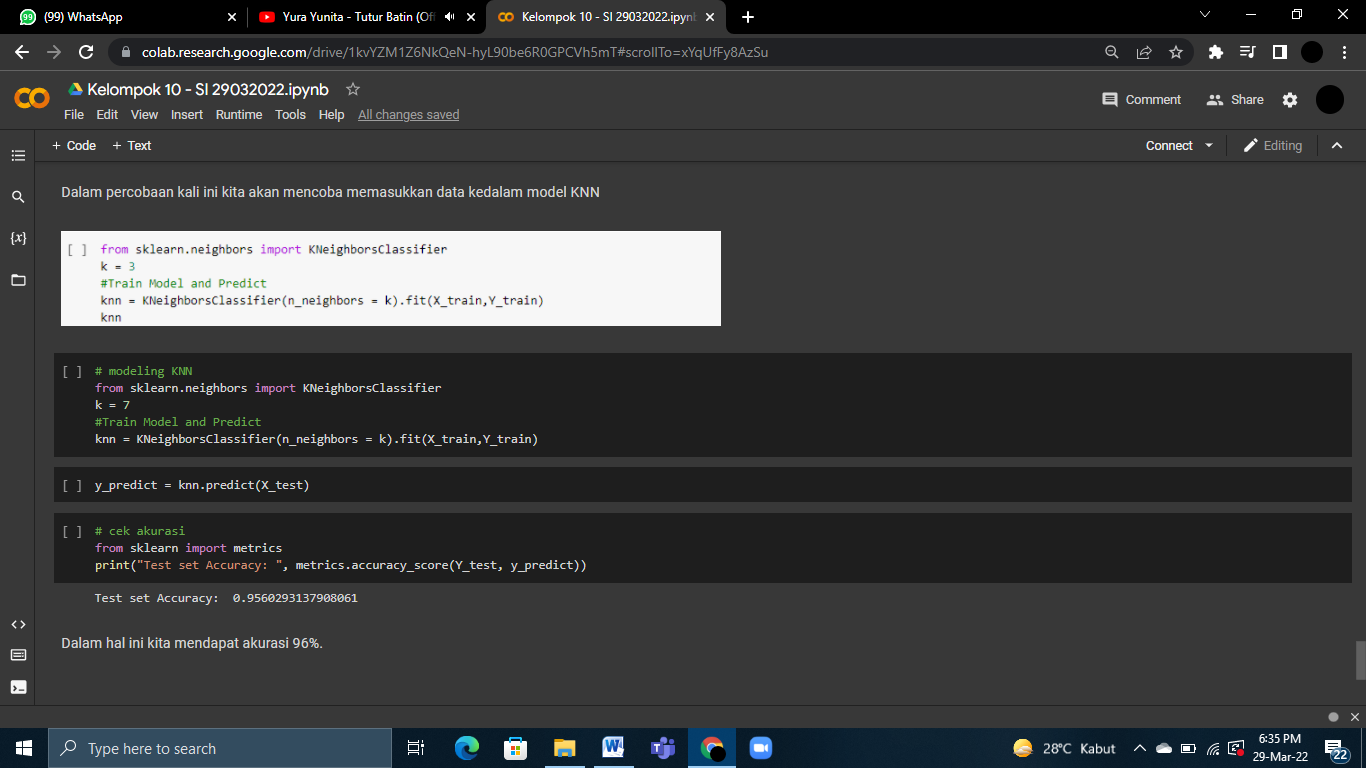


1. **Scale your Features**

Dalam sebuah dataset, biasanya nilai pada tiap fitur mempunyai range yang berbeda. Hal ini dapat menjadikan masalah untuk beberapa model machine learning. Untuk itu kita harus melakukan scaling agar semua nilai pada fitur memiliki range yang sama. Pada kasus ini kita menggunakan StandardScaler







Diperoleh nilai akurasi maksimal pada nilai k = 7 senilai 96%.