**LEMBAR SAMPUL PENGAJUAN TUGAS**

|  |  |
| --- | --- |
| **Nama Mahasiswa:** | **Javan Visman** |
| **Nomor Mahasiswa:** | **201855202042** |
| **Nama Mata Kuliah:** | **Data Mining** |
| **Nama Pengajar:** | **MUHAMMAD YUSUF, S.Kom., M.Kom.** |
| **Judul Tugas:** | UTS |
| **Penugasan Ke:** | 1 |
| **Tanggal Batas Pengumpulan:** | 29 Mei 2023 |

**PERNYATAAN**

Dengan ini saya menyatakan bahwa:

1. Tugas ini sepenuhnya adalah hasil pekerjaan saya sendiri. Selain yang tertera pada daftar pustaka, materi yang terkandung dalam tugas ini belum pernah dikirimkan untuk penilaian dalam tugas apa pun.
2. Saya telah memiliki salinan dari tugas ini dan jika diperlukan saya dapat menujukkan salinan dari tugas tersebut.
3. Saya telah memahami definisi dan konsekuensi dari plagiarisme.

|  |  |
| --- | --- |
| **Tanda Tangan Mahasiswa:** | **Tanggal:**  **Senin, 29 Mei 2023** |

Apa itu Preprocessing Data?

*data preprocessing* adalah teknik yang digunakan untuk mengubah data mentah dalam format yang berguna dan efisien. Inisiatif ini diperlukan karena data mentah seringkali tidak lengkap dan memiliki format yang tidak konsisten. Kualitas data sendiri memiliki korelasi langsung dengan keberhasilan setiap proyek yang melibatkan analisis data. *Preprocessing* sendiri melibatkan validasi dan imputasi data. Tujuan dari validasi adalah untuk menilai tingkat kelengkapan dan akurasi data yang tersaring.

Dalam *machine learning,* kegiatan ini sangat penting untuk memastikan bahwa [big data](https://glints.com/id/lowongan/apa-itu-big-data/#.YXulCp5Bzrc)sudah diformat dan informasi yang dikandungnya dapat ditafsirkan dan algoritma perusahaan.

Berikut adalah keempat tahap kerja *data preprocessing :*

1. ***Data cleaning***

Dalam tahap kerja ini, data dibersihkan melalui beberapa proses seperti mengisi nilai yang hilang, menghaluskan *noisy data*, dan menyelesaikan inkonsistensi yang ditemukan. Data juga bisa dibersihkan dengan dibagi menjadi segmen-segmen yang memiliki ukuran serupa lalu dihaluskan (*binning*). Kamu juga bisa menyesuaikannya dengan fungsi regresi linear atau berganda (*regression*), atau dengan mengelompokkannya ke dalam kelompok-kelompok data yang serupa (*grouping*).

1. ***Data integration***

Tahap kerja berikutnya dalam proses data preprocessing adalah data integration. Di sini, data dengan representasi yang berbeda disatukan dan semua konflik dalam di dalamnya diselesaikan. Tahap kerja satu ini merupakan proses lanjutan dari data cleaning dengan tujuan untuk membuat data lebih halus.

1. ***Data transformation***

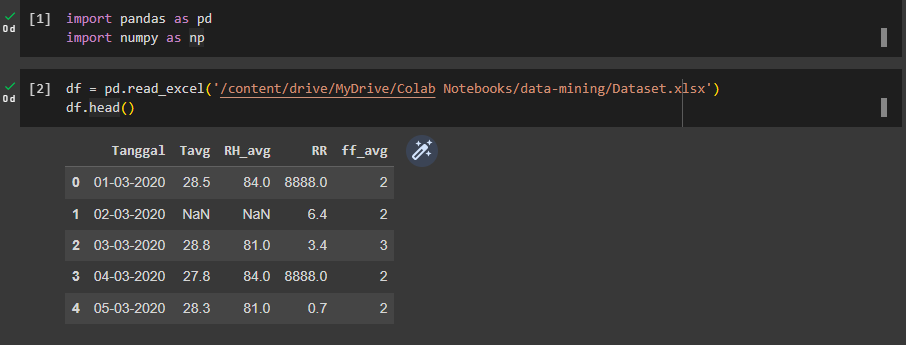
Data transformation adalah tahap kerja selanjutnya dalam proses data preprocessing. Pada tahap ini, data akan dinormalisasi dan digeneralisasikan. Normalisasi sendiri adalah sebuah proses di mana perusahaan memastikan bahwa tidak ada data yang berlebihan.Semua data akan disimpan dalam satu tempat dan semua dependensinya haruslah logis. Langkah ini juga diambil untuk mentransformasikan data ke dalam bentuk yang sesuai untuk proses mining.

1. ***Data reduction***

Tahap kerja terakhir dalam proses kerja data preprocessing adalah data reduction. Data mining adalah sebuah teknik yang digunakan untuk menangani data dalam jumlah yang besar. Saat bekerja dengan volume data yang besar, proses analisis akan menjadi lebih sulit. Nah, untuk mempermudah proses [data mining](https://glints.com/id/lowongan/apa-itu-data-mining/#.YXulG55Bzrc), kamu bisa menggunakan teknik data reduction. Sebab, menurut [Monkey Learn,](https://monkeylearn.com/blog/data-preprocessing/) inisiatif ini bisa meningkatkan efisiensi penyimpanan dan mengurangi representasi data dalam data warehouse.

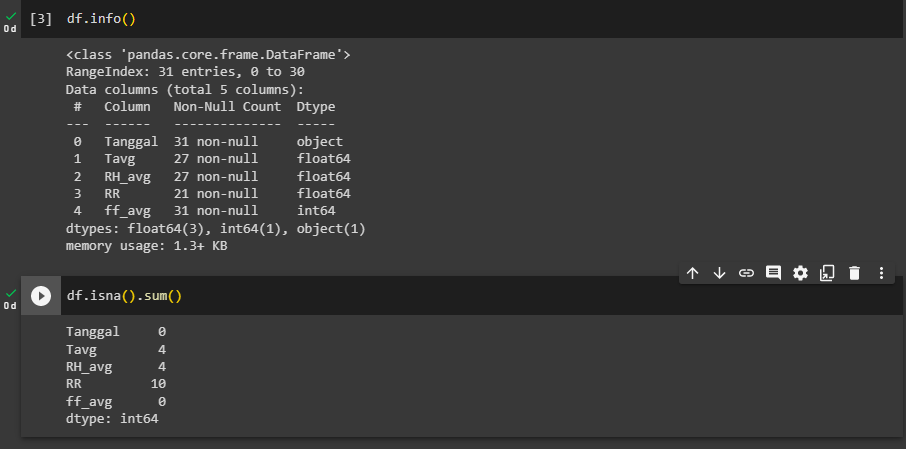
Berikut adalah mengolah Data baru menggunakan googlescollab :

**Import library dan dataset yang digunakan**

****

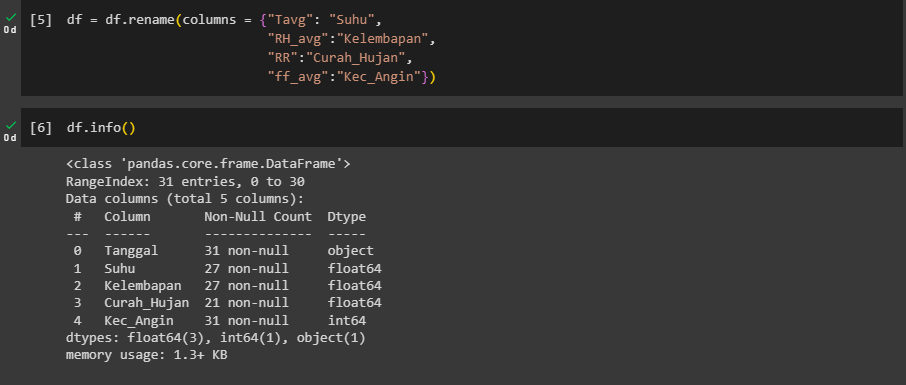
Sebelumnya kita import dahulu library yang akan dipakai seperti pandas dan numpy, kemudian akses dataset yang akan diolah. Disini dataset yang akan diolah telah saya upload pada G-Drive agar lebih mudah untuk pengaksesan datasetnya.

**Menampilkan info dan missing value pada data**



Selanjutnya kita tampilkan info pada dataset menggunakan perintah df.info() dan cek missing values atau nilai-nilai yang hilang pada dataframe/dataset kita menggunakan perintah df.isna().sum(). Dari sini dapat kita ketahui jumlah kolom dan baris hingga type data pada dataset kita dan juga jumlah missing values yang terdapat pada data kita.

**Mengubah nama atribut**

****

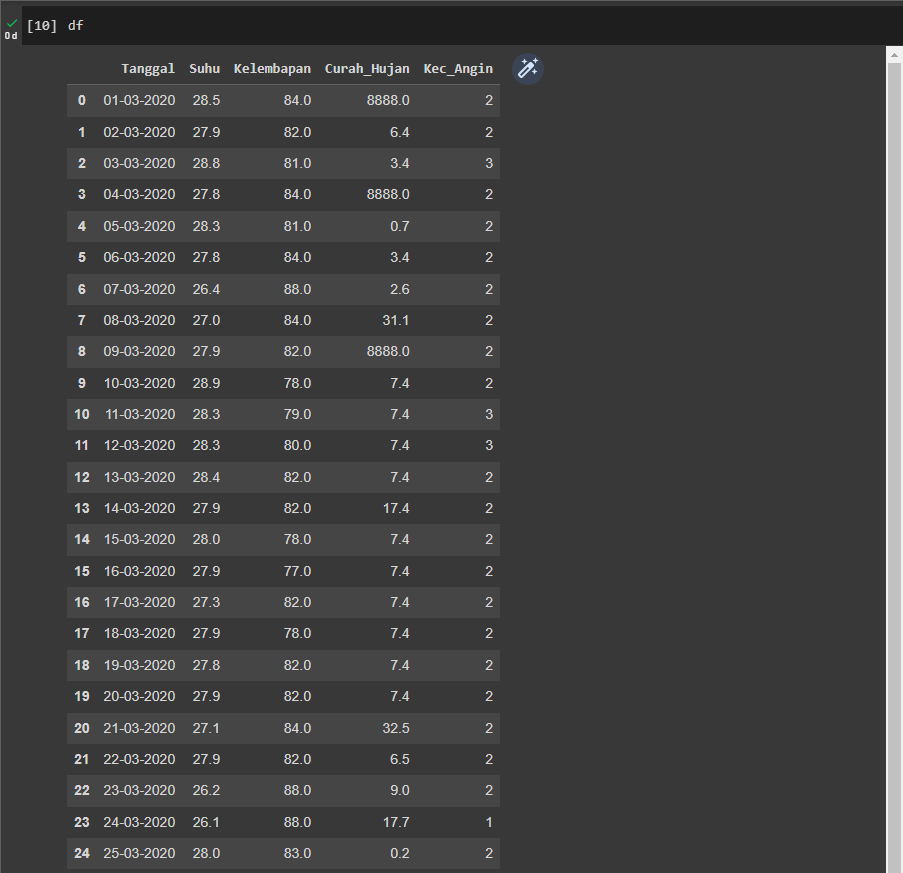
Berikutnya kita ubah dahulu nama atribut atau kolom pada datasetnya agar lebih mudah dipahami, dimana atribut yang kita ubah namanya adalah Tavg, RH\_avg, RR, ff\_avg menjadi Suhu, Kelembapan, Curah\_Hujan, Kec\_Angin.

**Menangani missing value**



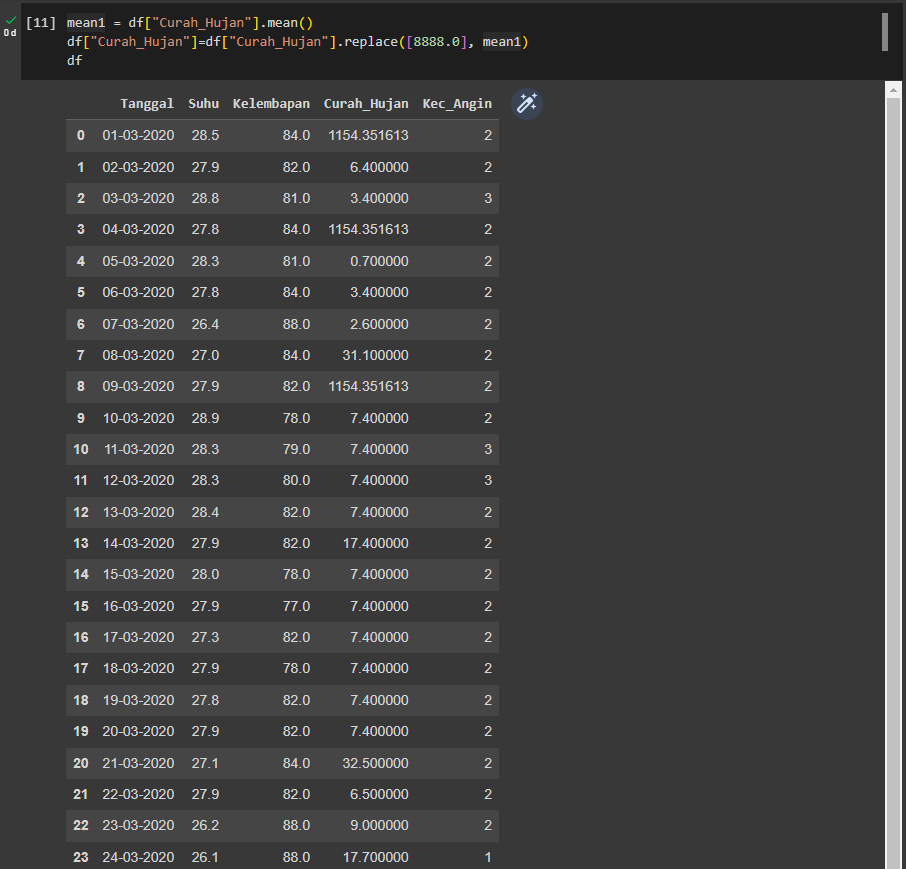
Pada tahap ini kita akan menangani missing values atau nilai-nilai yang hilang pada datasetnya yaitu 4 data Suhu, 4 data Kelembapan, dan 10 Data Curah\_hujan. Dan untuk mengisi data-data yang kosong tersebut kita akan menggunaka nilai median dari masing-masing atributnya.

**Menampilkan seluruh isi pada dataset**

****

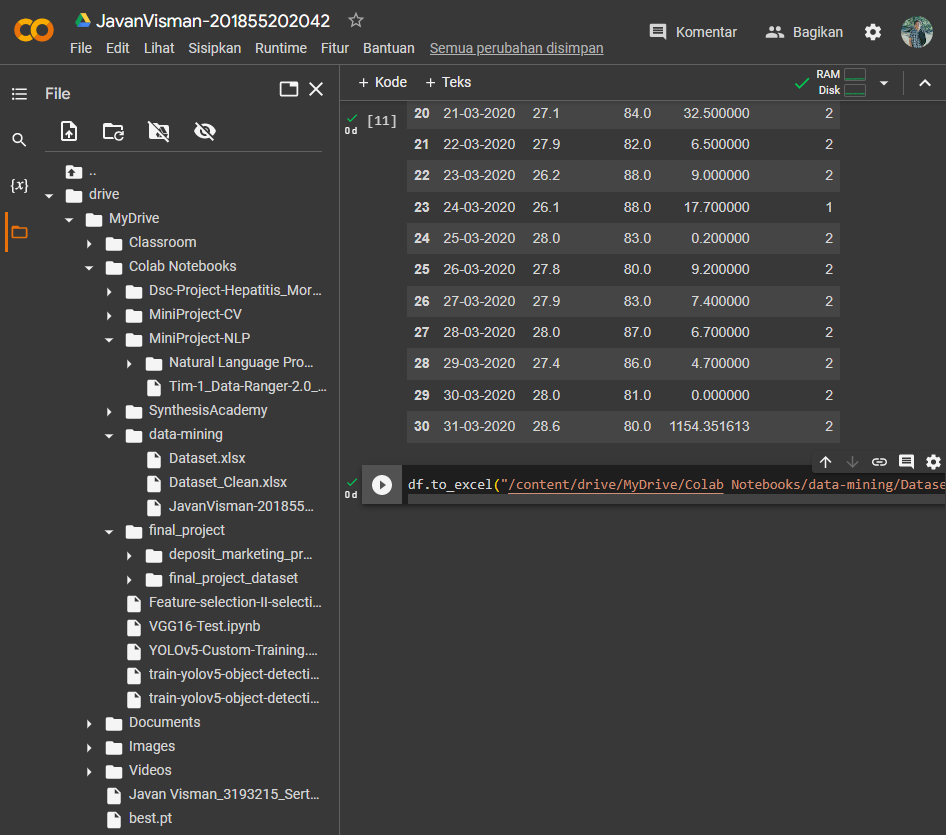
Disini dapat dilihat pada datanya terdapat nilai yang tidak beraturan dan terpaut jauh seperti 8888.0 yang mana ini merupakan data rusak atau outliers pada isi datasetnya yang juga akan bepengaruh bilah nantinya data ini akan kita gunakan.

**Menangani nilai 8888 yang merupakan outliers**



Untuk menangani nila tersebut kita akan menggunakan cara yang hampir sama dengan mengisi missing values sebelumnya, yaitu kita akan menggantikan nilai 8888 tersebut menjadi nilai meannya. Lalu cek kembali dataset yang telah olah tersebut.

**Menyimpan Dataset**



Sampai sini dataset yang sudah diolah tersebut sudah bersih dan lebih baik dari sebelumnya, selanjutnya simpan data tersebut dengan format yang diinginkan.