**PRAKTIKUM DATA WAREHOUSING DAN DATA MINING**

**MODUL 8**

**ALGORITMA NAÏVE BAYES**

****

**Disusun oleh:**

**Adinda Aulia Hapsari**

**L200220037**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS KOMUNIKASI DAN INFORMATIKA**

**UNIVERSITAS MUHAMMADIYAH SURAKARTA**

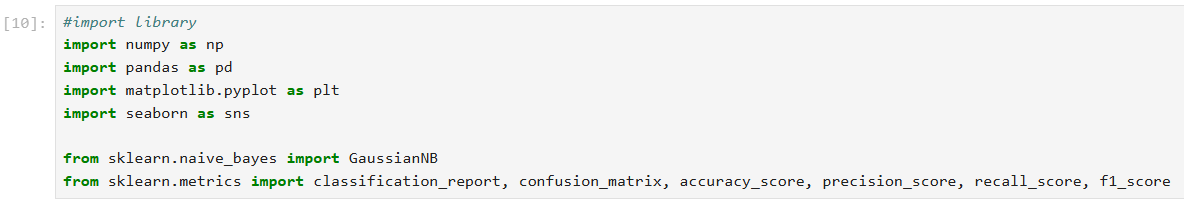
**TAHUN 2024**

|  |  |
| --- | --- |
| Setelah kegiatan selesai, lembar kerja ini dicetak (di-print) dan dikumpulkan ke asisten.  NIM : L200220037  Nama : Adinda Aulia Hapsari  Nama Asisten : Diva Halimah  Tanggal Praktikum : 22 November 2024 | (Diisi oleh Asisten)  Nilai Praktek :  Tanda Tangan : |

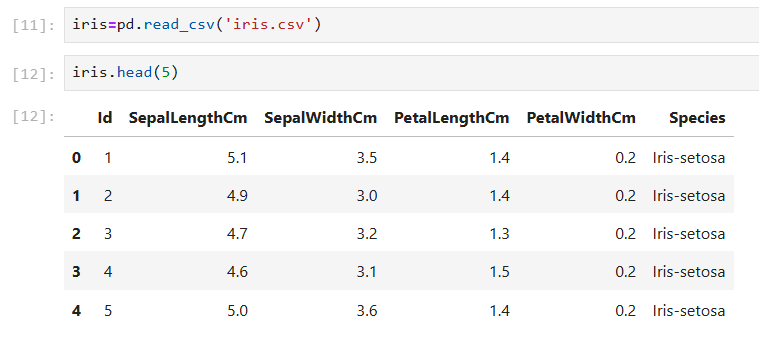
**KEGIATAN PRAKTIKUM**

Langkah-langkah menggunakan algoritma naïve bayes dengan bahasa pemrograman python adalah sebagai berikut:

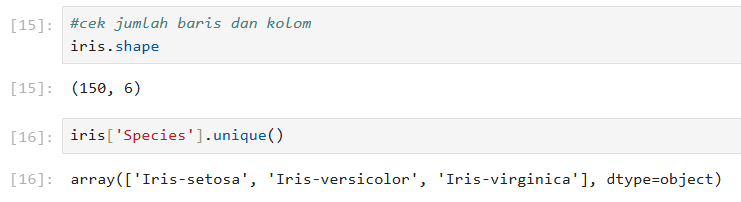
1. Bukalah Jupyter notebook yang sudah terinstall pada komputer yang digunakan.
2. Buatlah file baru pada jupyter notebook dengan mengklik menu New-Python 3, secara otomatis akan keluar window baru untuk melakukan coding pada jupyter nootebook. Pastikan file yang dibuat dengan dataset yang digunakan berapa dalam satu folder.
3. Pada baris pertama tuliskan beberapa library yang akan digunakan dalam praktikum. Ada beberapa library seperti numpy, pandas, matplotlib, seaborn dan lain sebagainya.



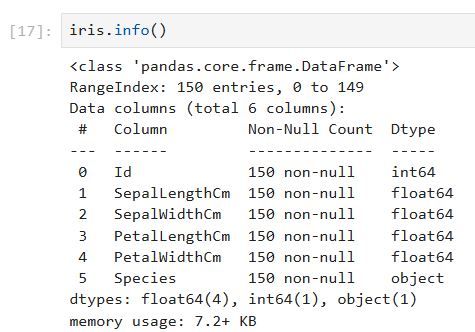
1. Setelah itu lakukan read dataset menggunakan library pandas, disini kita menggunakan dataset iris dengan ekstensi .csv. Kita akan melakukan beberapa hal untuk mengekplorasi data yang digunakan. Pertama kita akan melihat isi dataset yang dipakai menggunakan fungsi head, fungsi tersebut akan menampilkan isi data beserta atribut yang ada. Terdapat 6 atribut yang digunakan pada dataset tersebut yaitu Id, SepalLengthCm, SepalWidthCm, PetalLengthCM, PetalWidthCm dan Species yang nanti digunakan sebagai label klasifikasi.



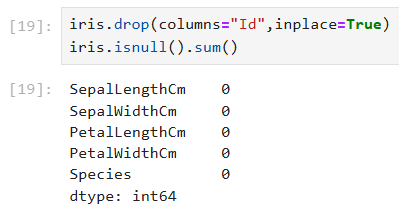
1. Kita juga bisa mengecek bentuk dari data yang digunakan menggunakan fungsi shape, bisa juga melihat value apa saja yang ada pada atribut Species.



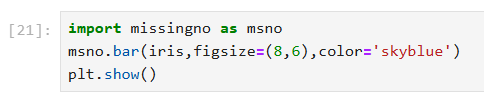
1. Hal lain yang harus kita lakukan adalah memeriksa apakah dataset memiliki nilai kosong (null) atau tidak, hal ini bisa dilakukan dengan menggunakan fungsi info(). Terlihat tidak ada data yang kosong pada semua atribut, ini menunjukan dataset yang digunakan sangat baik.

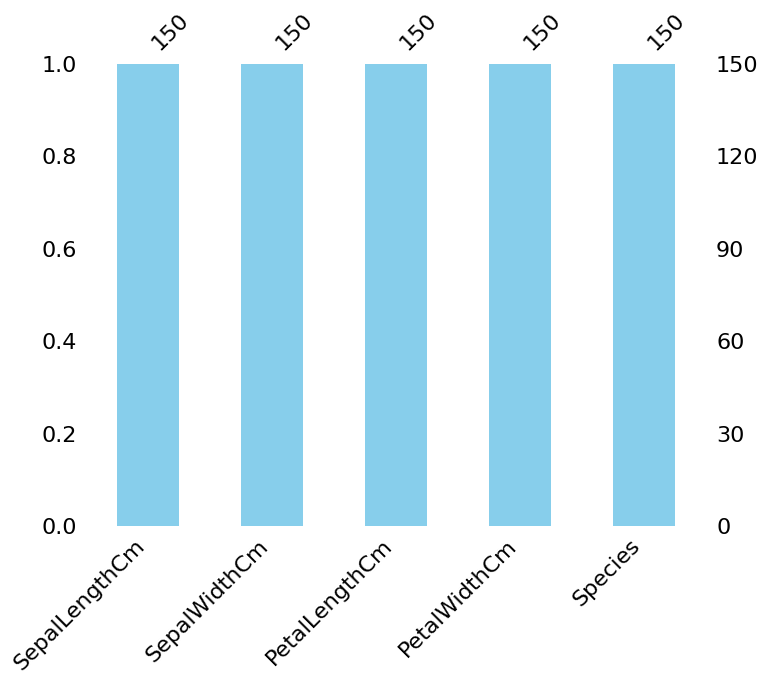


1. Dari semua atribut yang ada kita akan memakai lima fitur saja yaitu SepalLengthCm, SepalWidthCm, PetalLengthCM, PetalWidthCm dan Species. Sehingga atribut Id akan dihapus karena fitur tersebut tidak memiliki pengaruh dalam proses klasifikasi, penghapusan atribut bisa menggunakan fungsi drop.

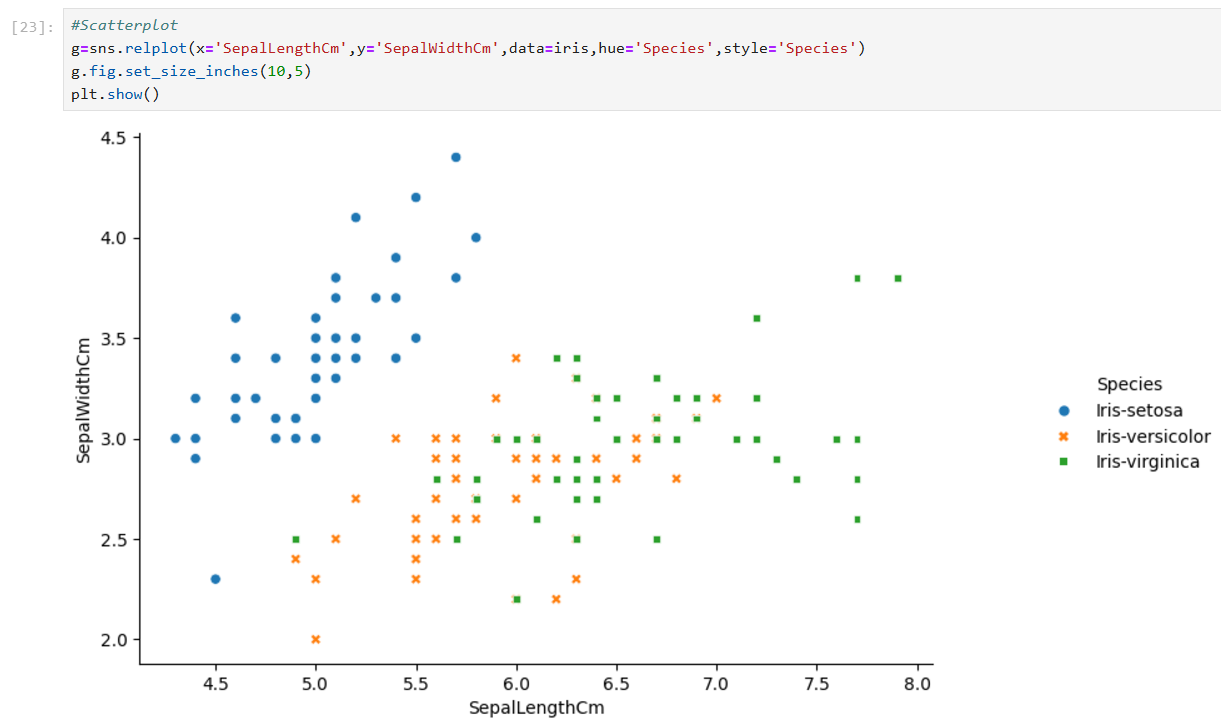


1. Kita juga bisa mengecek data yang kosong menggunakan grafik.

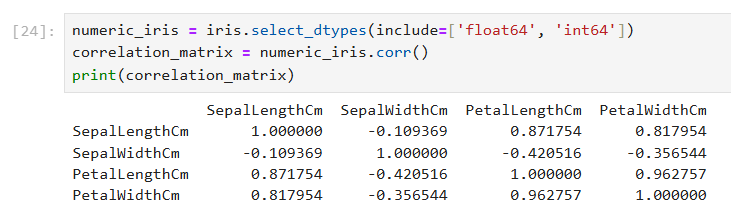


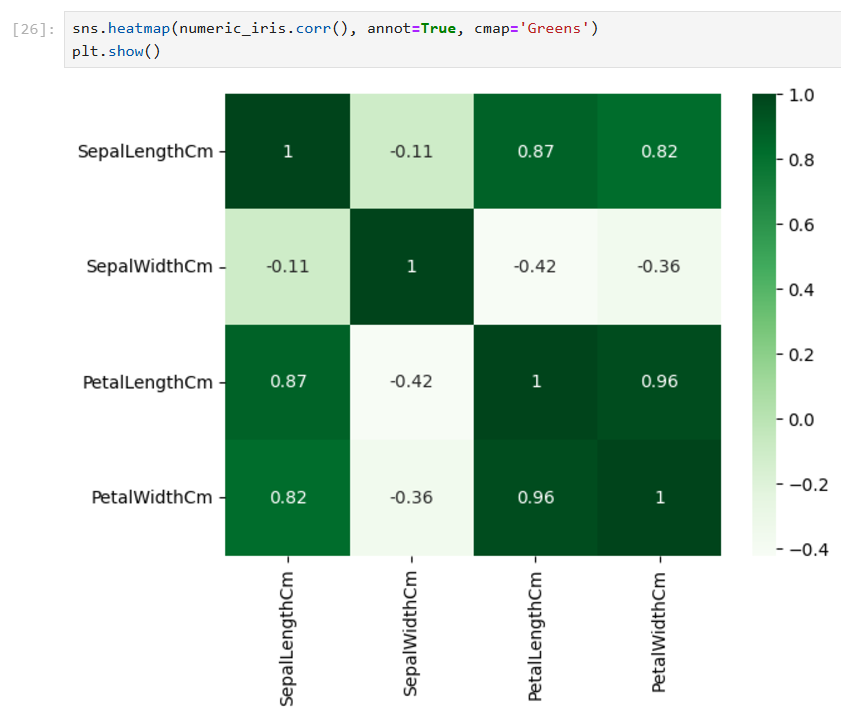


1. Proses selanjutnya adalah melakukan visualisasi data, proses ini sangat penting dilakukan untuk mengetahui sebaran data yang ada dalam dataset.



1. Kemudian kita bisa mencari korelasi pada setiap fitur yang digunakan untuk melakukan klasifikasi.



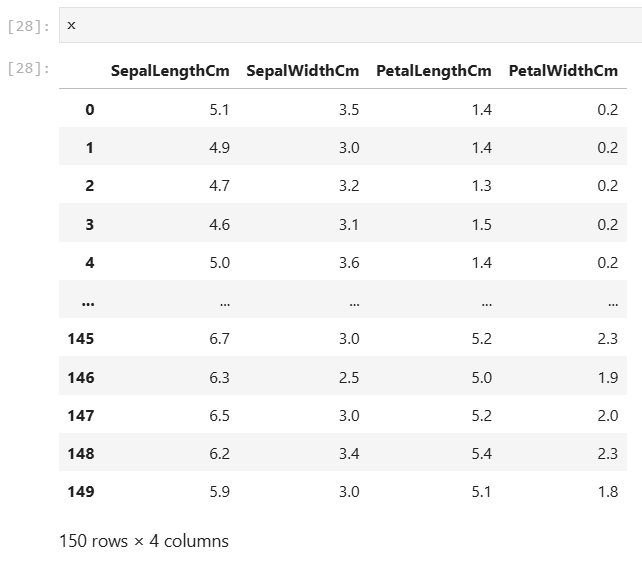


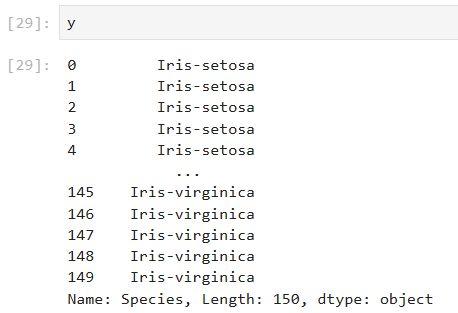
1. Hal penting lainnya yang harus dilakukan adalah memisahkan antara fitur dan label, atribut Species akan digunakan sebagai label dan atribut lainnya digunakna sebagai fitur pada proses klasifikasi.

**Data Preprocessing**

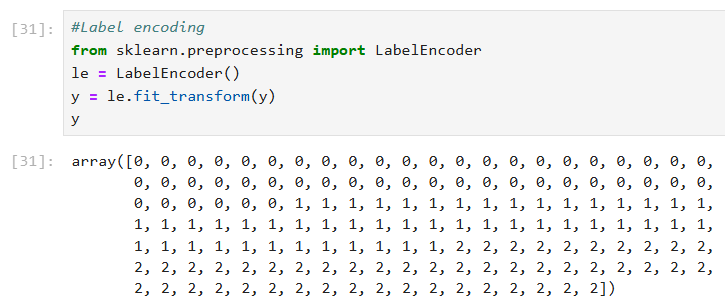
****

1. Kita bisa mengecek atribut yang sudah terpisah dengan menampilkan variabel X yang merupakan fitur dan variabel y sebagai label.

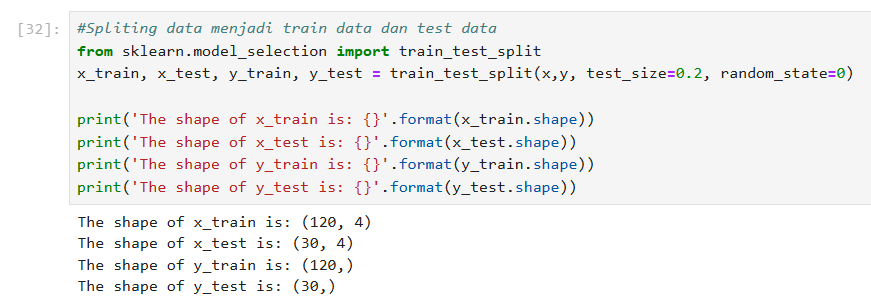




1. Pada dataset iris, label yang digunakan bukan merupakan data numerical, padahal mesin hanya bisa memahami angka, sehingga kita harus merubah isi dari label menjadi angka, proses ini dinamakan dengan encoding.

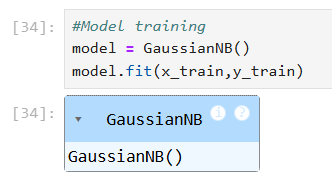


1. Sebelum memasuki proses klasifikasi data perlu dibagi menjadi dua bagian yaitu data training dan data testing. Data training akan digunakan untuk pembuatan model, sedangkan data testing digunakan untuk melakukan evaluasi pada model yang sudah dibuat, proses pembagian dataset ini menggunakan library sklearn. Terlihat pada gambar dataset terbagi menjadi 120 untuk data training dan 30 untuk data testing.

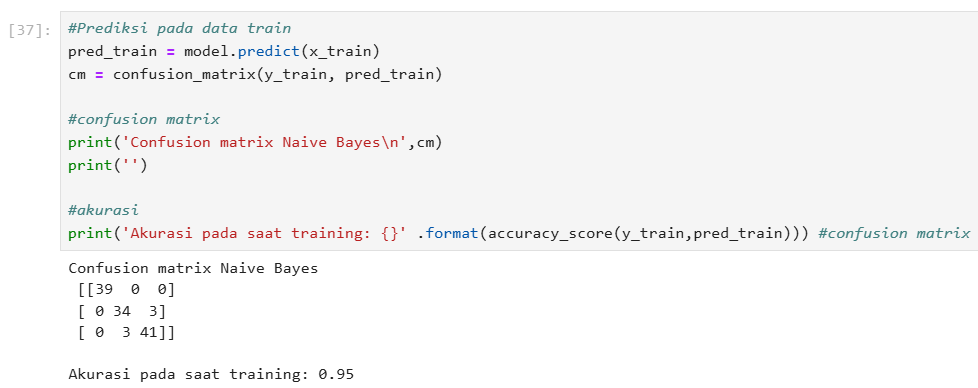


1. Tahap selanjutnya adalah pembuatan model menggunakan algoritma Naive Bayes. Kita bisa memanggilnya dari library sklearn dan menaruhnya pada variabel model.

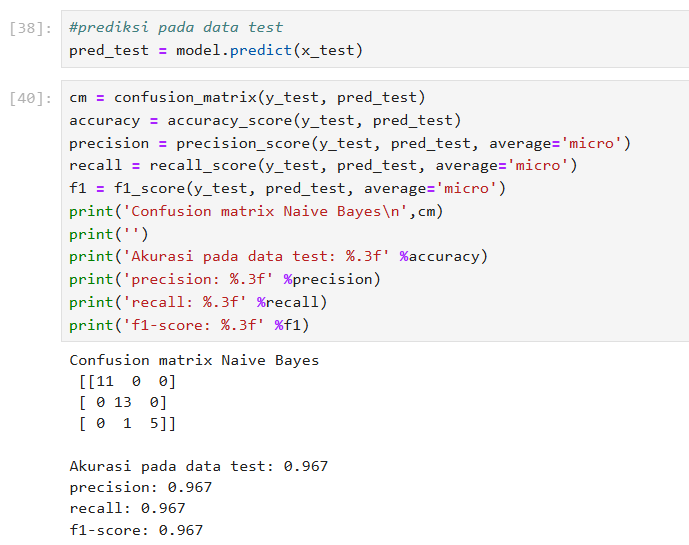
**Membuat Model**



1. Kita lakukan training model menggunakan algoritma naive bayes pada data training. Disini kita mendapatkan akurasi pada saat training memperoleh skor 95%.



1. Langkah terakhir adalah menguji model yang dibuat dengan data testing yang sudah disiapkan. Pada saat evaluasi model kita mendapatkan nilai yang sama pada akurasi, presisi, recall dan f1 score, yaitu sebesar 96,7%.

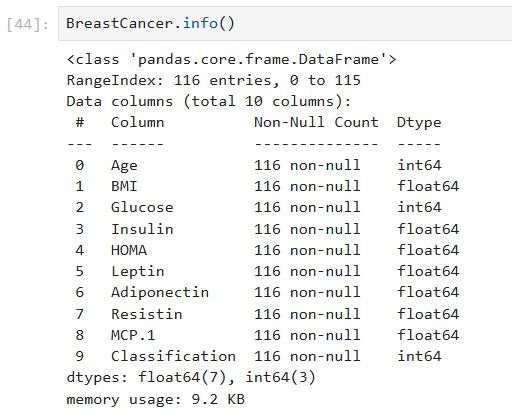


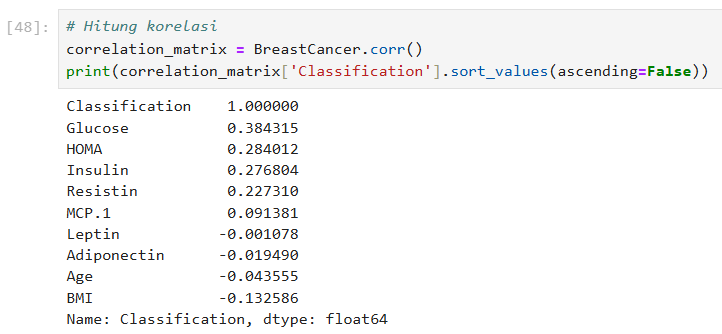
**TUGAS**

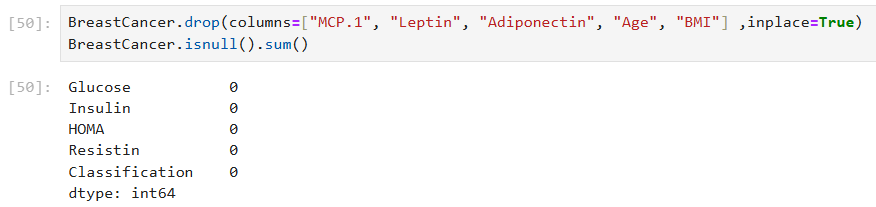
1. Buatlah ekplorasi data dari dataset Breast-Cancer.csv!



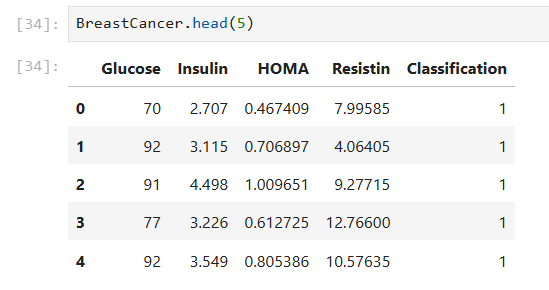
1. Cari 4 fitur yang memiliki pengaruh paling besar terhadap proses klasifikasi!

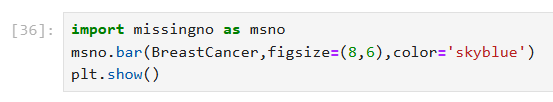


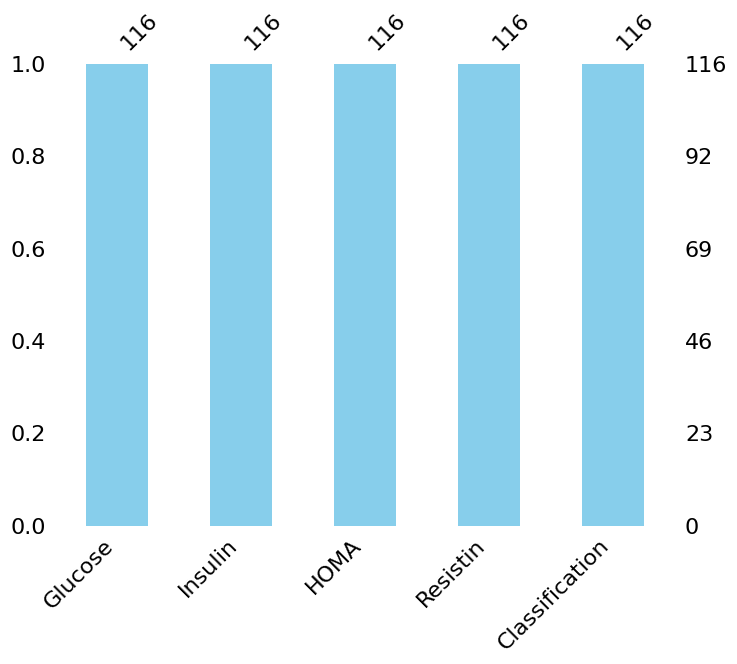


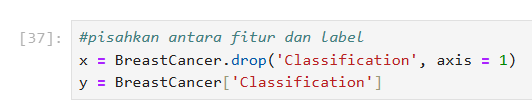


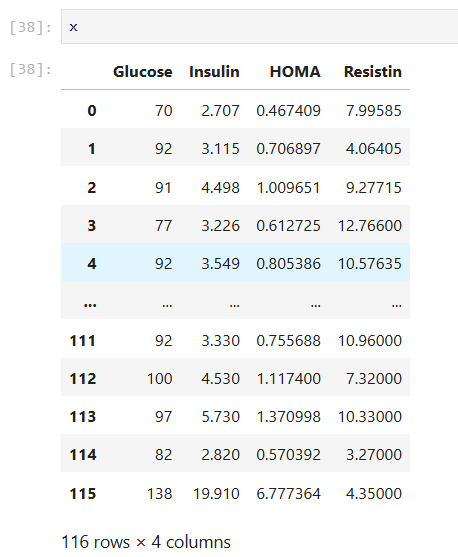
1. Gunakan 4 fitur yang sudah dipilih untuk melakukan klasifikasi!

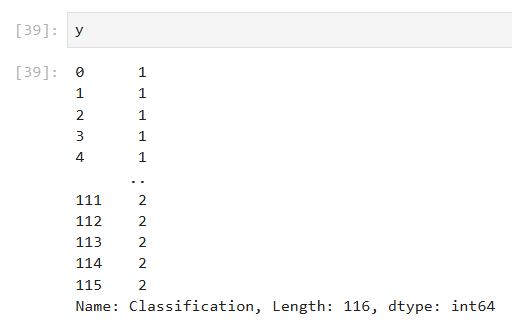


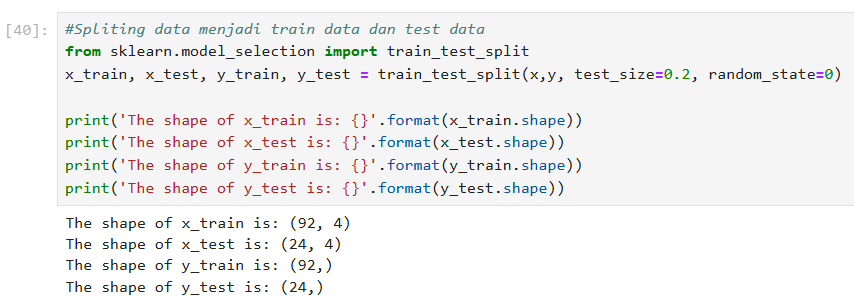












1. Buatlah model untuk algoritma Naive Bayes dan hitunglah nilai akurasi, presisi, recall dan F1-score!

