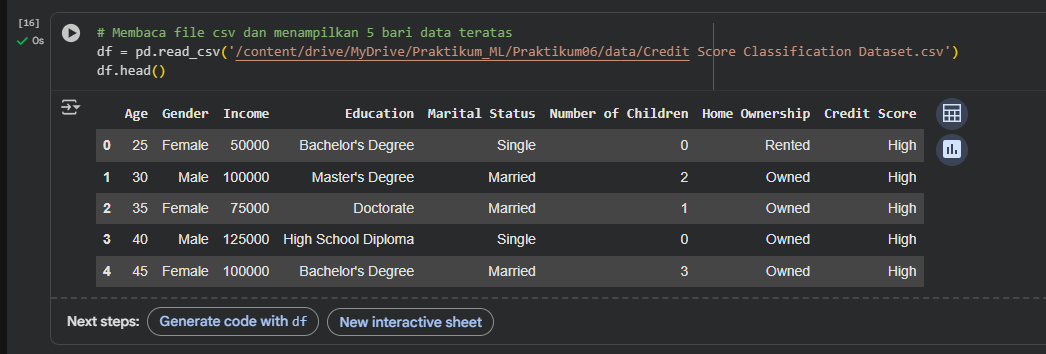
**Tugas 5: Praktikum Mandiri 5 Machine Learning**

**Aria Kristallinacht Sundanis - 0110222076**

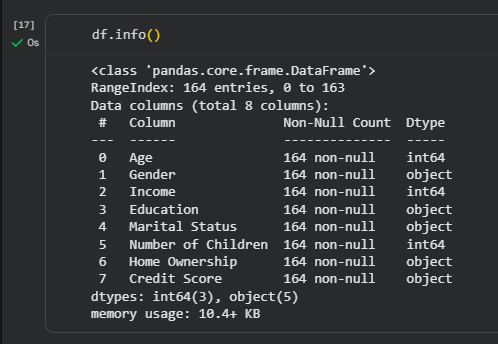
Teknik Informatika, STT Terpadu Nurul Fikri, Depok

E-mail: [aria22076ti@student.nurulfikri.ac.id](mailto:aria22076ti@student.nurulfikri.ac.id)

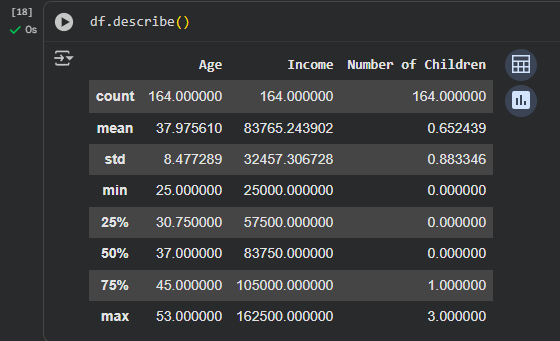
**Praktikum Mandiri 5**



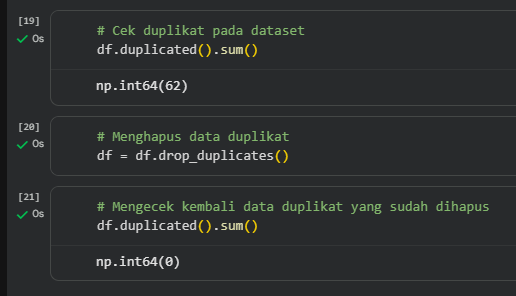
Pada tahap ini membaca dataset Credit Score Classification.csv yang sudah disimpan di Google Drive lalu dipanggil melalui sebuah variabel path. Setelah itu, file tersebut dibaca dengan fungsi pd.read\_csv() dan hasilnya disimpan ke dalam sebuah DataFrame dengan nama df, untuk memastikan data berhasil dimuat, digunakan perintah df.head() yang menampilkan lima baris pertama dari dataset. Secara keseluruhan, dataset ini berisi 164 baris dan 8 kolom yang merepresentasikan data penyewaan sepeda.



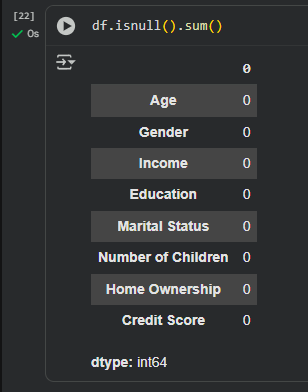
Perintah df.info() digunakan untuk menampilkan informasi umum mengenai struktur dataset. Dari hasil tersebut terlihat bahwa dataset memiliki 164 baris dan 8 kolom, dengan semua kolom berisi data lengkap, tidak ada missing value dan tipe data yang digunakan adalah int64 dan object.



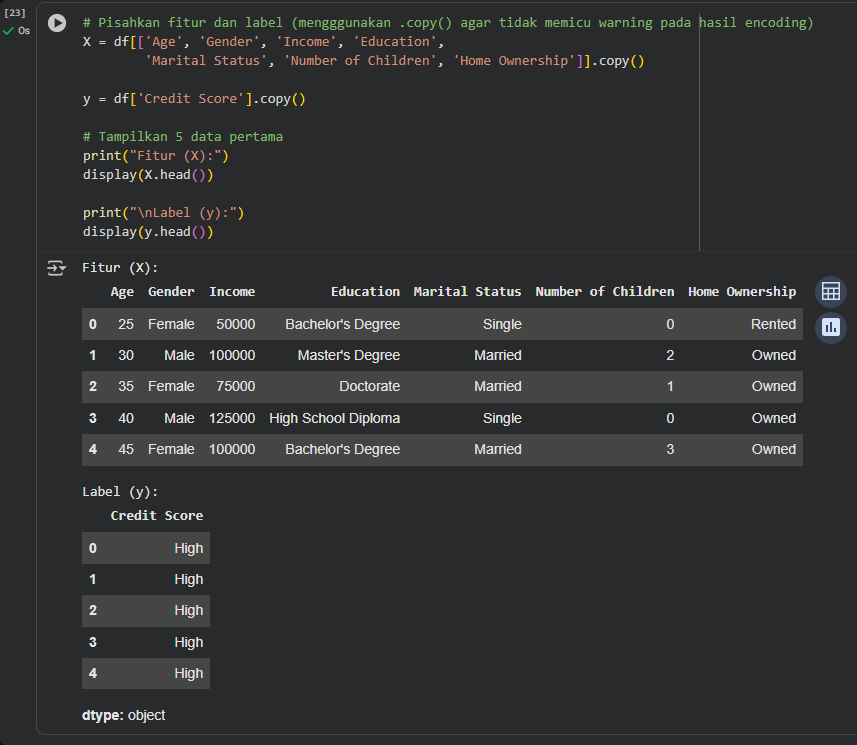
Perintah df.describe() digunakan untuk menampilkan statistik deskriptif dari setiap kolom numerik dalam dataset. Bisa dilihat bahwa dataset tersebut memiliki jumlah data 164 pada tiap kolom. Rata-rata usia responden adalah sekitar 38 tahun dengan rentang antara 25 hingga 53 tahun. Pendapatan responden bervariasi cukup besar, dengan rata-rata sekitar 83.765 dan nilai minimum 25.000 hingga maksimum 162.500. Sementara itu, jumlah anak memiliki rata-rata 0,65 dengan nilai maksimum 3, yang menunjukkan bahwa sebagian besar responden belum memiliki anak.



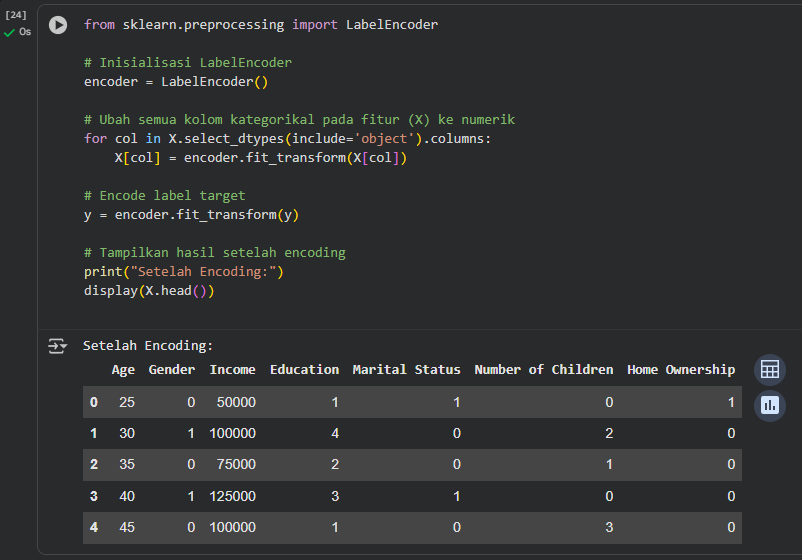
Lalu tahap selanjutnya yaitu mencek dan menghapus duplikat pada dataset, bisa dilihat dataset Credit Score Classification.csv memiliki 64 duplikat. Karna ada duplikasi yang lumayan banyak, maka harus dihapus dengan perintah df.drop\_duplicates() supaya hasil analisis lebih akurat dan tidak bias.



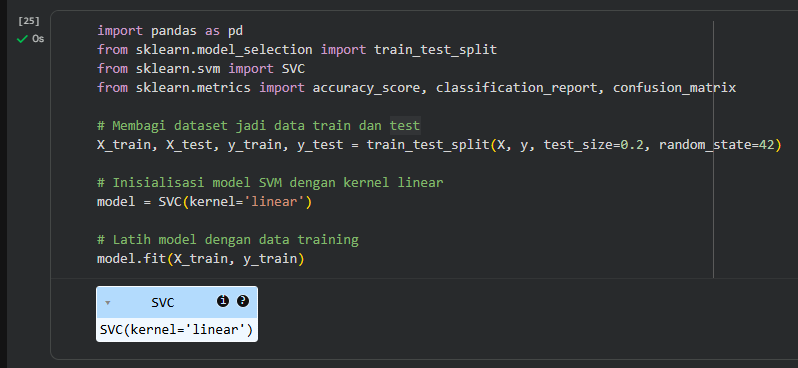
Perintah df.isnull().sum() digunakan untuk mengecek mising value, hasilnya menunjukkan bahwa seluruh kolom yang ada di dataset Credit Score Classification.csv tidak ada missing value.



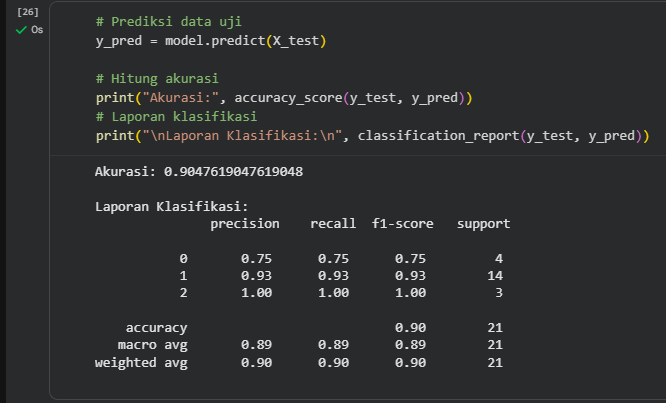
Setelah itu dilakukan pemisahan antara fitur (X) dan label (y) dari dataset. Fitur (X) berisi variabel-variabel yang akan digunakan sebagai input untuk model, yaitu Age, Gender, Income, Education, Marital Status, Number of Children, dan Home Ownership. Sedangkan label (y) berisi variabel target yang ingin diprediksi, yaitu Credit Score. Penggunaan .copy() bertujuan agar tidak muncul peringatan (warning) ketika dilakukan proses encoding atau manipulasi data selanjutnya. Dari hasil yang ditampilkan, terlihat bahwa lima data pertama memiliki variasi karakteristik demografis, sementara label Credit Score pada contoh awal semuanya bernilai “High”.



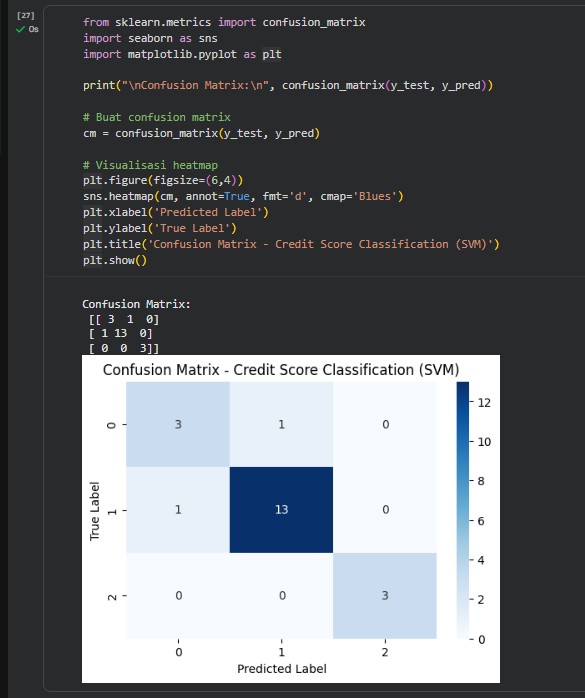
Pada tahap ini dilakukan proses encoding untuk mengubah data kategorikal menjadi bentuk numerik menggunakan LabelEncoder dari library scikit-learn. Proses ini penting karena sebagian besar algoritma machine learning hanya dapat memproses data dalam format angka. Kode melakukan iterasi pada setiap kolom bertipe objek (kategorikal) dalam fitur X dan mengonversinya ke nilai numerik. Selain itu, label target y juga diubah menjadi angka agar dapat digunakan dalam proses pelatihan model. Hasilnya, setiap kategori seperti Gender, Education, Marital Status, dan Home Ownership kini direpresentasikan dalam bentuk kode numerik, misalnya “Female” menjadi 0 dan “Male” menjadi 1.



Pada tahap ini dilakukan proses pembagian dataset menjadi data latih (training set) dan data uji (testing set) menggunakan fungsi train\_test\_split, dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Setelah itu, diinisialisasi model Support Vector Machine (SVM) dengan kernel linear, yang berarti model akan mencari garis pemisah lurus (hyperplane) terbaik untuk memisahkan kelas pada data. Selanjutnya, model dilatih menggunakan data training (X\_train dan y\_train) agar dapat mempelajari pola hubungan antara fitur-fitur input dengan label Credit Score.



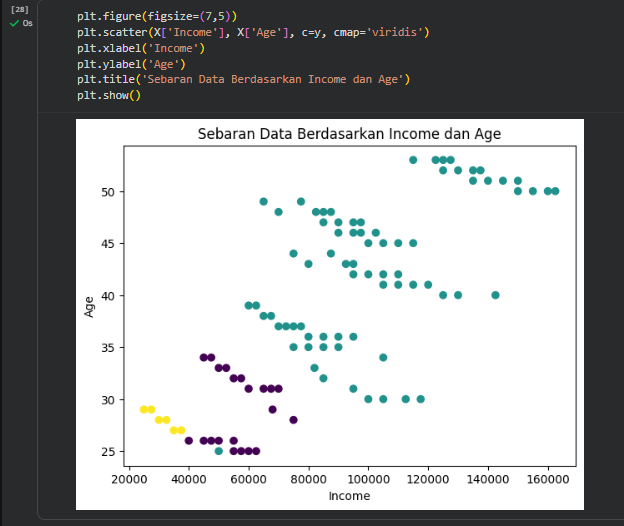
Setelah model SVM (Support Vector Machine) selesai dilatih, dilakukan proses prediksi terhadap data uji (X\_test) untuk mengevaluasi performa model. Hasilnya menunjukkan bahwa model memiliki akurasi sebesar 90,47%, yang berarti model mampu memprediksi nilai Credit Score dengan tingkat ketepatan sekitar 90%. Berdasarkan laporan klasifikasi, setiap kelas memiliki nilai precision, recall, dan f1-score yang tinggi, terutama pada kelas 2 dengan skor sempurna (1.00). Hal ini menunjukkan bahwa model mampu mengenali pola pada data dengan baik dan memiliki performa yang cukup stabil di semua kelas.



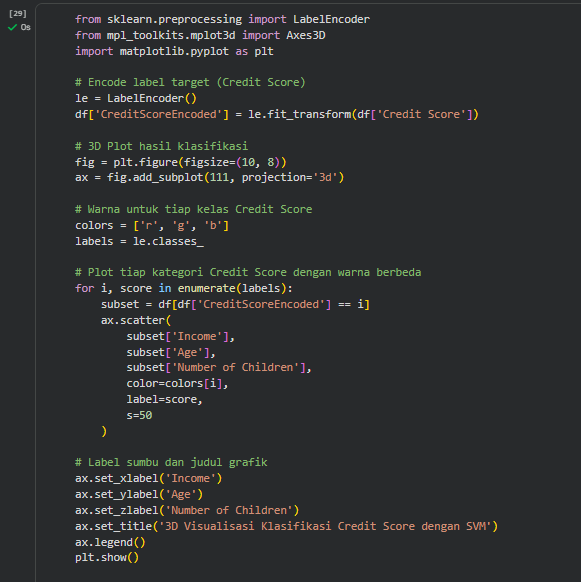
Setelah data selesai dilatih, lalu membuat dan menampilkan confusion matrix dari hasil prediksi model SVM. Confusion matrix berfungsi untuk melihat seberapa baik model mengklasifikasikan setiap kelas dibandingkan dengan label aslinya. Dari hasil yang ditampilkan, dapat dilihat bahwa:

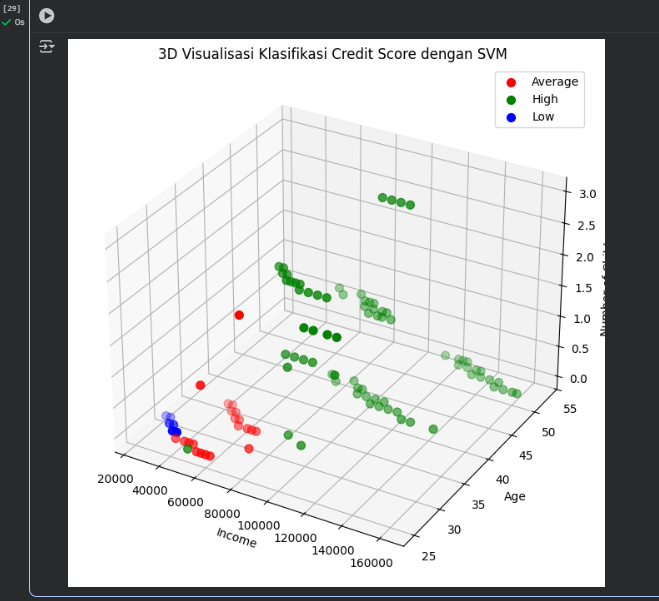
1. Kelas 0 (baris pertama) memiliki 3 data yang benar diklasifikasikan sebagai kelas 0, dan 1 data yang salah diklasifikasikan sebagai kelas 1.
2. Kelas 1 (baris kedua) memiliki 13 data yang diklasifikasikan dengan benar, serta 1 data yang salah diklasifikasikan sebagai kelas 0.
3. Kelas 2 (baris ketiga) semuanya diklasifikasikan dengan benar sebanyak 3 data.

Visualisasi menggunakan heatmap membantu memperjelas distribusi prediksi model. Warna biru tua menunjukkan jumlah prediksi yang benar paling banyak (kelas 1), sedangkan warna biru muda menandakan adanya kesalahan klasifikasi pada kelas lain. Secara keseluruhan, model SVM menunjukkan performa yang sangat baik karena sebagian besar data berhasil diklasifikasikan dengan benar.



Setelah itu membuat visualisasi sebaran data (scatter plot) antara variabel Income dan Age, dengan warna titik mewakili nilai Credit Score (y). Grafik ini memperlihatkan hubungan antara tingkat pendapatan dan usia responden. Dari hasil visualisasi, terlihat bahwa responden dengan usia dan pendapatan yang lebih tinggi (bagian kanan atas grafik) cenderung memiliki satu kategori Credit Score tertentu (warna hijau kebiruan), sedangkan responden muda dengan pendapatan lebih rendah (bagian kiri bawah) cenderung berada pada kategori Credit Score lain (warna ungu atau kuning). Pola ini menunjukkan bahwa faktor usia dan pendapatan memiliki pengaruh yang cukup jelas terhadap klasifikasi Credit Score pada dataset tersebut.





Setelah membuat visualisasi 2D, disini saya juga membuat visualisasi 3D yang menampilkan hasil klasifikasi Credit Score berdasarkan tiga variabel utama, yaitu Income, Age, dan Number of Children. Setiap titik pada grafik merepresentasikan satu individu dalam dataset, sementara warna menunjukkan kategori Credit Score yang berbeda merah untuk Average, hijau untuk High, dan biru untuk Low. Dari grafik terlihat bahwa individu dengan pendapatan tinggi dan usia lebih matang cenderung memiliki Credit Score tinggi (warna hijau), sedangkan individu dengan pendapatan rendah dan usia muda lebih banyak berada pada kategori Low (biru) dan Average (merah). Visualisasi ini membantu memperjelas pola hubungan antara faktor demografis dengan tingkat Credit Score, serta menunjukkan pemisahan antar kelas yang cukup jelas sesuai hasil klasifikasi model SVM.

Link Kaggle : <https://www.kaggle.com/datasets/sujithmandala/credit-score-classification-dataset>

Link GitHub :