Nama	: Ratika Dwi Anggraini
Kelas	: TK-44-06
NIM	: 1103201250
Dataset	: Titanic Dataset (https://www.kaggle.com/datasets/brendan45774/test-file)
Pengolahan Data	: Classification
Model	: XGBoost

Script Video YouTube

Assalamualaikum wr. wb., Saya Ratika Dwi Anggraini dari kelas TK4406 dengan NIM: 1103201250 Prodi S1 Teknik Komputer Universitas Telkom. Saya akan menjelaskan mengenai pengolahan data secara classification dengan model XGBoost pada Titanic Dataset. Pada penjelasan ini, saya menggunakan tools chatgpt, google collab, dan google drive.

Hal pertama yang perlu dilakukan yaitu mengupload Titanic Dataset ke google drive. Jadi nantinya kita akan menyambungkan google collab dengan drive sehingga kita bisa akses dataset melalui drive tersebut. Untuk file yang kita upload Namanya yaitu Titanic Dataset.

Step selanjutnya yaitu:

1. mencari library apa saja yang dibutuhkan menggunakan chatgpt dengan prompt **"Beri kode untuk mengimport library pandas, numpy, matpolib.pyplot, seaborn, sklearn,dan xgboost untuk mengklasifikasi dataset dan menvisualisasikan data serta EDA"**. Kemudian kita copy paste dan jalankan kode tersebut.

   Nah disini ada (pd)Pandas: digunakan untuk manipulasi dan analisis data. (np)NumPy : Operasi dasar aritmatika python yabng dapat digunakan untuk bekerja dengan data numerik yang besar. (plt)Matplotlib : Digunakan untuk membuat visualisasi data seperti grafik dan plot. (sns)Seaborn : Memanfaatkan matplotlib untuk membuat grafik statistik yang menarik dan informatif. (sklearn)Scikit-learn : Digunakan untuk analisis data dan evaluasi model. (xgb)XGBoost : metode ensemble learning yang membangun serangkaian model lemah dan menggabungkan hasil prediksi mereka untuk menciptakan model yang lebih kuat secara keseluruhan.

2. Kemudian kita menghubungkan google collab dengan google drive. Kita cari menggunakan prompt chatgpt **“Beri kode untuk mengimport dataset menggunakan google colab dari drive”** Okay kemudian kita copy paste dan jalankan kode tersebut.

   Sebelumnya, mount google drive yaitu Mengimpor modul drive dari library google.colab kemudian menggunakan fungsi mount untuk mengaitkan Google Drive dengan sesi Colab. Kemudian disini pd.read_csv digunakan untuk membaca data dari file CSV. Path file CSV yang digunakan berada di direktori Google Drive (/content/drive/MyDrive/Colab Notebooks/Titanic_Dataset.csv).

   Hasil running menunjukkan Mounted at /content/drive yang berarti Proses mounting (mengaitkan) Google Drive dengan Google Colab telah berhasil dilakukan atau Google Colab sekarang memiliki akses ke file dan direktori yang ada di Google Drive.
3. Selanjutnya lakukan EDA. Exploratory Data Analysis (EDA) adalah suatu pendekatan dalam analisis data yang bertujuan untuk memahami karakteristik dan struktur data secara visual dan statistic. Kita cari menggunakan chatgpt prompt **“Beri kode untuk menampilkan seluruh baris pertama pada dataset”**

   kemudian bisa dilihat disini terdapat Kode print(data.head()) yang digunakan untuk menampilkan beberapa baris pertama dari dataset yaitu memberikan gambaran awal tentang struktur data, jenis variabel, dan nilai-nilai dalam dataset tersebut.

   Hasil running menampilkan informasi dataset, yaitu terdapat kolom PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, dan Embarked. Nah disini kita mencoba memprediksi apakah penumpang selamat atau tidak berdasarkan kolom survived.
4. Setelah melihat seluruh baris pertama dataset, kita lanjut menampilkan informasi dataset dengan prompt **“Beri kode untuk menampilkan info dataset”** lalu copy paste dan jalankan kode tersebut.

   Disini tertulis ‘print(data.info())` yang digunakan untuk mendapatkan informasi secara rinci mengenai dataset yang menghasilkan output seperti berikut (diem dulu sambal nunjukin output 5 detik). Terdapat beberapa informasi yang dapat kita ambil yaitu ternyata dataset ini memiliki 418 baris data dan 12 kolom yang tidak memiliki nilai *null*. Untuk tipe data yang terdapat pada dataset ini yaitu integer, float, dan object.
5. Tahap selanjutnya yaitu melakukan data visualization. Data visualization sendiri merupakan Penggunaan grafik dan visualisasi untuk memahami dan menganalisis data. Tujuannya

mengidentifikasi data yang tidak biasa (outlier), mengeksplorasi fitur, evaluasi model, dan komunikasi hasil dengan lebih efektif. Beberapa contoh visualisasi data yaitu scatter plots, heatmaps, dan bar chart. Sekarang kita cari menggunakan chatgpt prompt **“Beri kode untuk visualisasi EDA untuk mengetahui distribusi selamat dan tidak selamat berdasarkan Survived yg terdapat pada dataset”** kemudian copy paste dan jalankan kode

kode tersebut menghasilkan plot count yang menunjukkan distribusi antara penumpang yang selamat (Survived=1) dan tidak selamat (Survived=0) dalam dataset. Plot ini memberikan gambaran visual tentang seberapa seimbang atau tidak seimbangnya jumlah penumpang yang selamat dan tidak selamat.

Nah dari hasil running terlihat grafik batang dengan Jumlah penumpang yang selamat direpresentasikan oleh nilai x=1,sedangkan yang tidak selamat direpresentasikan oleh nilai x=0.

6. Setelah melihat grafik distribusi selamat dan tidak selamat, kita lanjut untuk memulai persiapan training data. Pertama-tama kita atur nilai yang hilang dengan mengisi menggunakan metode forward fill. Promptnya yaitu **"Beri kode untuk mengisi nilai yang hilang pada dataset menggunakan metode forward fill"** lalu copy paste dan jalankan kodenya. Dapat dilihat bahwa kode tersebut menggunakan metode `fillna` pada objek DataFrame `data` untuk mengisi nilai-nilai yang hilang (missing values) dalam dataset. Hal ini dapat membantu mengatasi masalah nilai yang hilang dalam analisis data, terutama jika nilai yang hilang dapat diestimasi atau diasumsikan dari nilai sebelumnya dalam dataset.
7. Selanjutnya kita cari menggunakan chatgpt prompt **“Beri kode untuk menerapkan one-hot encoding pada variabel kategorikal dalam dataset yang menghasilkan kategori yang lebih rendah dari setiap variabel kategorikal dihapus”**, kemudian seperti sebelumnya kita copy paste.

   Dari kode ini menghasilkan dataset baru yang telah diubah dengan one-hot encoding untuk kolom-kolom tertentu. Setiap kategori dalam kolom-kolom tersebut akan menjadi kolom baru dengan nilai biner (0 atau 1) yang menunjukkan keberadaan atau ketiadaan kategori tersebut. Sekarang kita jalankan kode tersebut.

   Bisa kita lihat bahwa kolom tersebut diubah menjadi bentuk baru dengan nilai biner (0 atau 1). Sebagai contoh, Sex diganti dengan dua kolom: Sex_male (menunjukkan apakahpenumpang adalah pria atau bukan) dan Sex_female (menunjukkan apakah

penumpangadalah wanita atau bukan). Kemudian Setelah one-hot encoding, dataset jadi memiliki total 864 kolom, yang melibatkan pembuatan banyak kolom baru yang masing-masing merepresentasikan keberadaan atau ketiadaan kategori tertentu dalam kolom asli.

8. Sebelum masuk ke training data, kita pisahkan terlebih dahulu fitur dan label dari variabel 'survived' dengan prompt **"Beri kode untuk memisahkan feature dan label dari variabel 'Survived' pada dataset kemudian membagi dataset menjadi data train dan data test"** lalu kita copy paste dan jalankan kode tersebut ke sini. Kode tersebut menjelaskan bahwa X berisi semua kolom dari dataset 'data' kecuali 'Survived' karena menggunakan metode 'drop' untuk menghapus kolom 'Survived'. Lalu Y menjadi variabel target yang berisi kolom 'Survived' dari dataset 'data'. Terakhir yaitu membagi data latih dan data uji menggunakan train_test_split dari scikit-learn.
9. Selanjutnya kita training data, Training data memberikan informasi yang diperlukan agar model dapat belajar dan disesuaikan untuk memahami karakteristik umum dari data. Training Data juga membantu model untuk menghasilkan prediksi yang akurat dan dapat diterapkan pada situasi yang belum pernah dilihat sebelumnya. Kita cari menggunakan chatgpt prompt **“Beri kode untuk menginisialisasikan dan melatih model menggunakan XGBoost”**, kemudian copy paste.

   Kode tersebut digunakan untuk membuat, melatih, dan menguji model machine learning menggunakan algoritma XGBoost. Selanjutnya kita jalankan dan dapat kita lihat bahwa pada hasil running terdapat model yang telah dilatih dan prediksi yang dihasilkan oleh model pada data uji.
10. Selanjutnya kita evaluasi data menggunakan chatgpt prompt sebelumya yaitu **“Beri kode untuk menginisialisasikan dan melatih model menggunakan XGBoost”**. Evaluasi ini dilakukan untuk mengukur seberapa baik model dapat memprediksi hasil yang benar pada data baru. Hasil evaluasi membantu memilih model terbaik, mengoptimalkan parameter, dan membuat keputusan berbasis data untuk tujuan bisnis.

    Kode dari chatgpt tersebut digunakan untuk menghitung dan mencetak akurasi dari model klasifikasi berdasarkan prediksi yang telah dilakukan terhadap data uji. Kemudian kita jalankan kode tersebut dan mendapatkan Accuracy: 100.00%. Hasil tersebut menunjukkan bahwa model klasifikasi yang telah dilatih memiliki akurasi 100% pada data uji. Akurasi sebesar 100% berarti bahwa model berhasil memprediksi dengan benar semua label kelas pada

data uji. Namun, perlu dicatat bahwa akurasi 100% bisa menjadi indikator bahwa ada sesuatu yang tidak beres. Beberapa poin untuk dipertimbangkan seperti overfitting dan ukuran dataset.

11. Setelah kita melatih model menggunakan XGBoost, kita lanjutkan untuk melatih menggunakan confusion matrix pada dataset dengan prompt **"Beri kode untuk melatih menggunakan confusion matrix pada dataset"** lalu kita copy paste dan jalankan kodenya. Dapat dilihat pada kode berikut bahwa cm berisikan fungsi confusion_matrix untuk menghasilkan confusion matrix, lalu terdapat label sebenarnya yaitu 'y_test' dan label prediksi yaitu 'y_pred'. Kita buat gambar berukuran 8x6 inci menggunakan 'plt.figure(figsize=(8,6))' dan kita definisikan heatmap. Terakhir, kita tampuilkan plot heatmap menggunakan 'plt_show()' dan terdapat hasilnya sebagai berikut. Setelah mengetahui grafiknya, kita buat laporan klasifikasi yang telah kita lakukan.
12. Setelah melakukan klasifikasi, maka kita melaporkan hasil klasifikasi tersebut dengan cara mencari chatgpt prompt **“Beri kode untuk report klasifikasi yang telah dilakukan”.** Kemudian kita copy paste dan jalankan kodenya.

    Hasil running yang diperoleh menunjukkan bahwa model memiliki performa yang sangat baik pada data uji dengan akurasi 100%. Semua metrik evaluasi seperti presisi, recall, dan F1-score memiliki nilai maksimum (1.00) untuk setiap kelas. Meskipun hasil ini terlihat sangat baik, perlu diingat bahwa kesimpulan ini dapat mencerminkan overfitting jika model terlalu "memorori" data latih dan tidak dapat menggeneralisasi dengan baik pada data baru. Evaluasi lebih lanjut dan validasi menggunakan dataset yang lebih besar dan beragam dapat memberikan pemahaman yang lebih baik.
13. Langkah terakhir yaitu mengevaluasi menggunakan data dummy untuk menguji data yang baru. Pertama kita persiapkan data dummy lalu masukkan prompt "Buatlah kode untuk new data input dengan input: (data dummy)" ke dalam chatgpt lalu copy paste dan jalankan kode tersebut. Setelah itu, pastikan kolom-kolmnya sesuai dengan X menggunakan prompt "Pastikan kolom-kolomnya sesuai dengan X_train.columns" lalu copy paste dan jalankan kodenya. Pada kode tersebut kita sesuaikan kolom data input baru (new_data) dengan X_train.columns lalu simpan hasil prediksi dalam variabel 'prediction' dan tampilkan hasil prediksinya. Ternyata outputnya menunjukkan bahwa dengan data input baru, modelnya memprediksi bahwa penumpang tersebut tidak selamat yang ditunjukkan oleh angka 0 (Mendefinisikan Not Survived).