LAPORAN PRAKTIKUM  
IF3270 PEMBELAJARAN MESIN

****

Oleh:

Eugene Yap Jin Quan 13521074

Dewana Gustavus Haraka Otang 13521173

**PROGRAM STUDI TEKNIK INFORMATIKA**

**SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA**

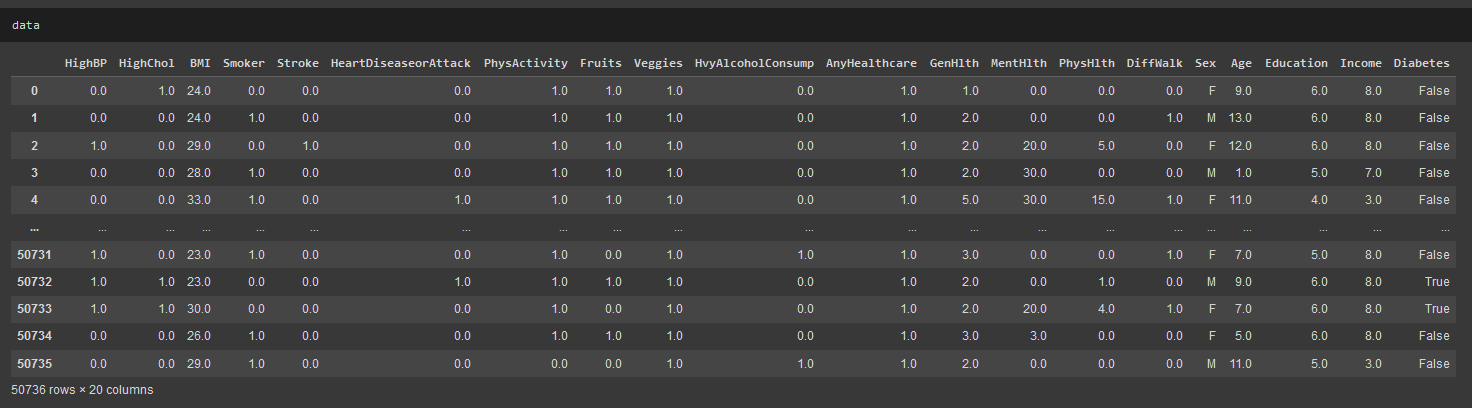
**INSTITUT TEKNOLOGI BANDUNG**

**2024**

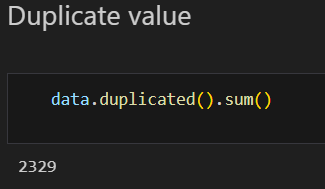
# HASIL ANALISIS DATA

Hasil analisis data adalah sebagai berikut.

* Terdapat 50736 baris, dan 20 kolom.



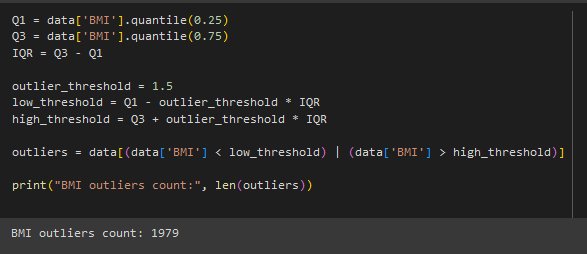
* Kolom “BMI” merupakan kolom numerik.
* Kolom “Sex” merupakan kolom kategorikal bernilai ‘M’ dan ‘F’.
* Kolom selain “BMI” dan “Sex” bernilai boolean ataupun skala.
* Kolom yang bernilai skala memiliki ukuran skala yang berbeda-beda (misal “MentHlth” berskala 0-30, “GenHlth” berskala 1-5).
* Ditemukan 2329 baris data duplikat pada dataset.



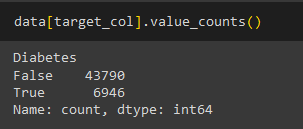
* Tidak ditemukan nilai missing value pada dataset



* Perhitungan outlier pada kolom “BMI” menggunakan nilai IQR menunjukkan bahwa terdapat baris outlier sebanyak 1979 baris.



* Nilai-nilai pada kolom target menunjukkan bahwa dataset bersifat imbalanced



# PENANGANAN DARI HASIL ANALISIS DATA

Berdasarkan hasil analisis data, penanganan yang perlu dilakukan adalah sebagai berikut.

* Encoding untuk mengubah nilai teks pada kolom “Sex” menjadi boolean numerik dengan cara labeling.
* Pembuangan baris duplikat.
* Pembuangan baris yang mengandung outlier.
* Resampling (diprioritaskan menggunakan cara oversampling) untuk menangani dataset imbalanced
* Normalisasi skala pada kolom-kolom atribut.

Penanganan tidak dilakukan untuk missing values karena dataset bersifat lengkap.

# JUSTIFIKASI TEKNIK-TEKNIK YANG DIPILIH

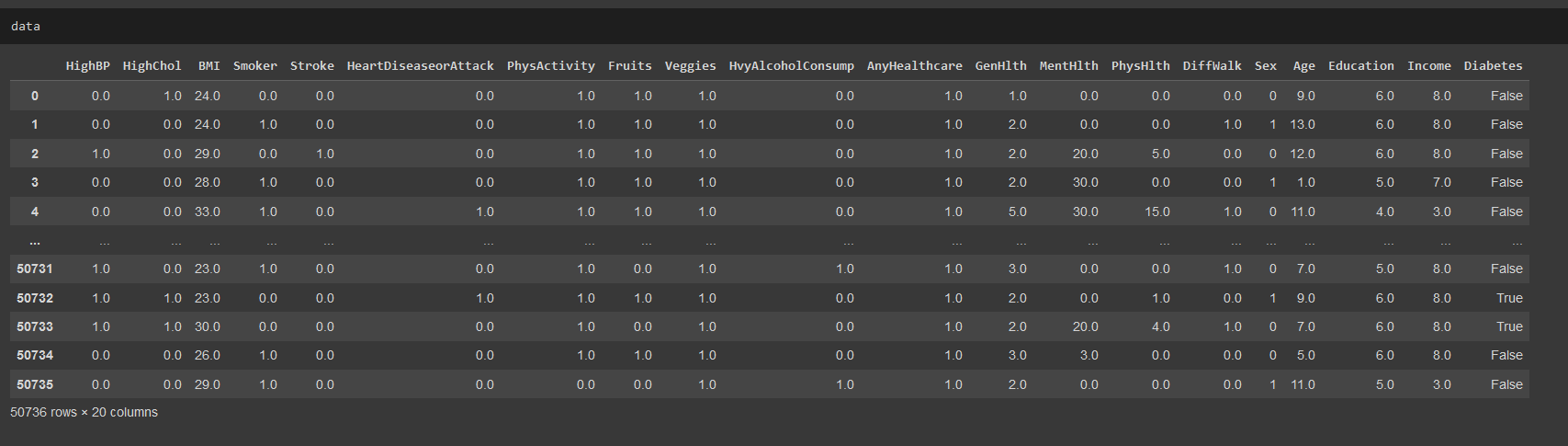
Berikut adalah justifikasi terhadap teknik-teknik penanganan data yang dipilih.

* Encoding label pada kolom “Sex” diperlukan untuk memasukkan data pada implementasi model yang disediakan pada library Python.
* Baris duplikat perlu dibuang agar performa model tidak menurun.
* Baris outlier dapat ditangani dengan cara membuangnya agar tidak mengacaukan statistik pada kolom.
* Resampling dengan cara oversampling lebih baik ketimbang undersampling karena perbandingan jumlah kelas minoritas jauh dibanding dengan jumlah kelas mayoritas (6946/43790). Penanganan dengan cara undersampling akan menyebabkan data loss pada kelas mayoritas.
* Normalisasi pada atribut-atribut dataset diperlukan agar setiap kolom pada dataset memiliki bobot pembelajaran yang sejenis.

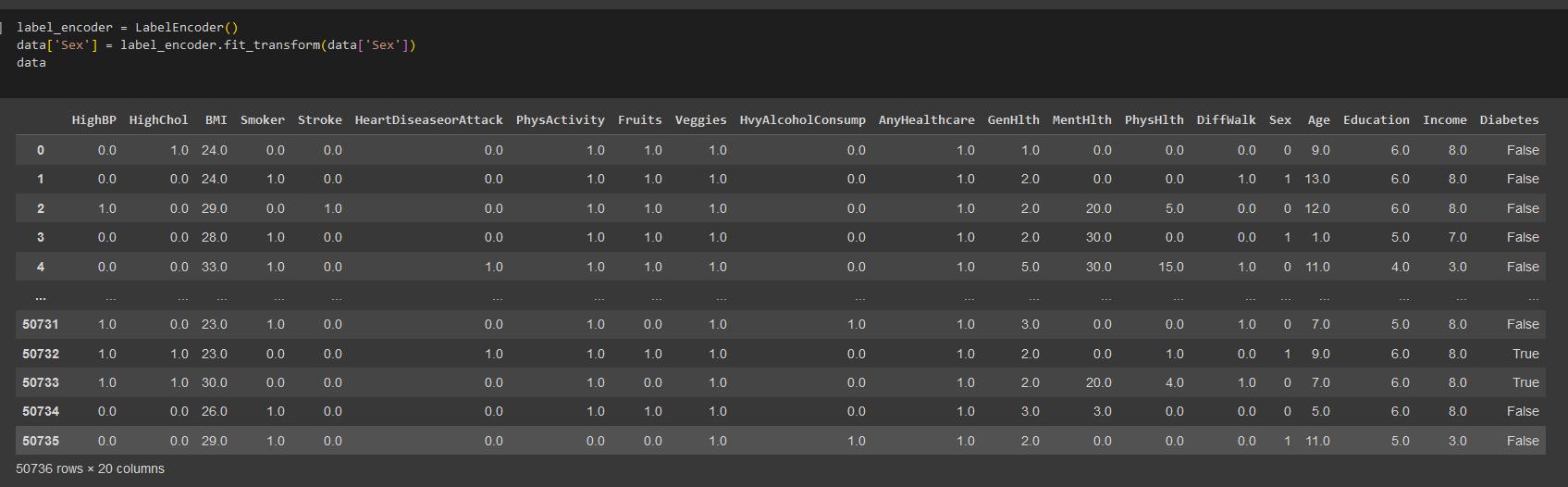
# PERUBAHAN YANG DILAKUKAN PADA JAWABAN POIN 1—5

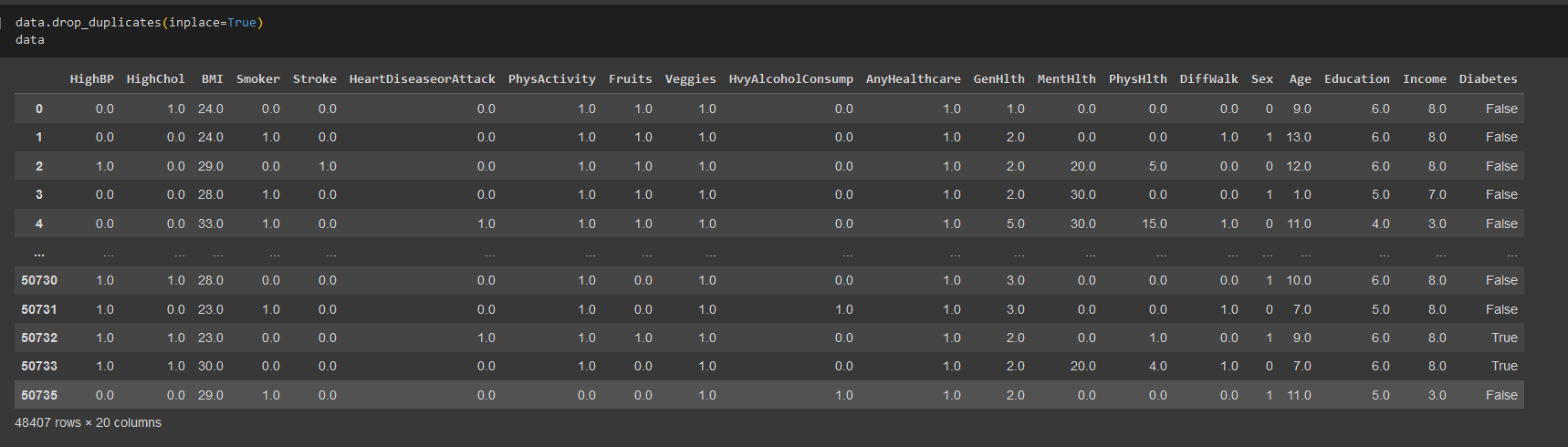
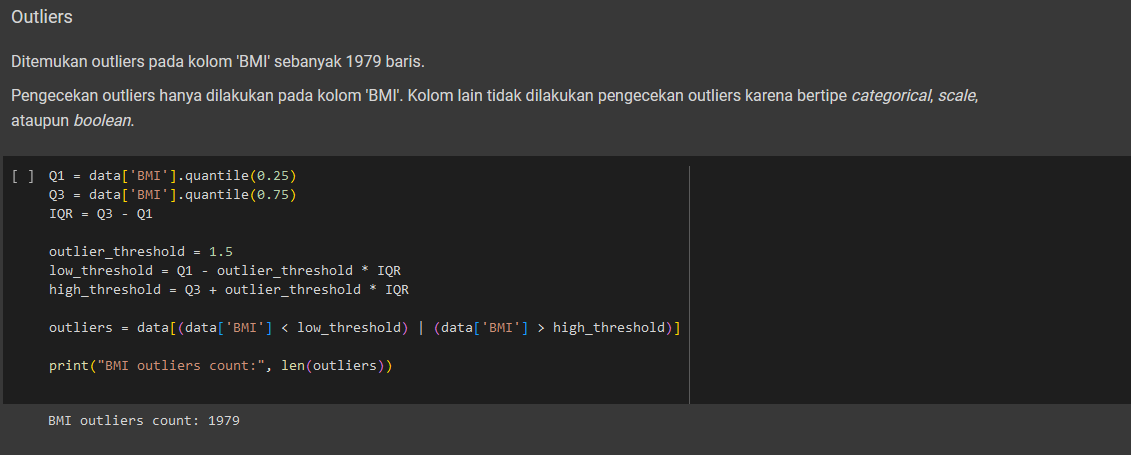
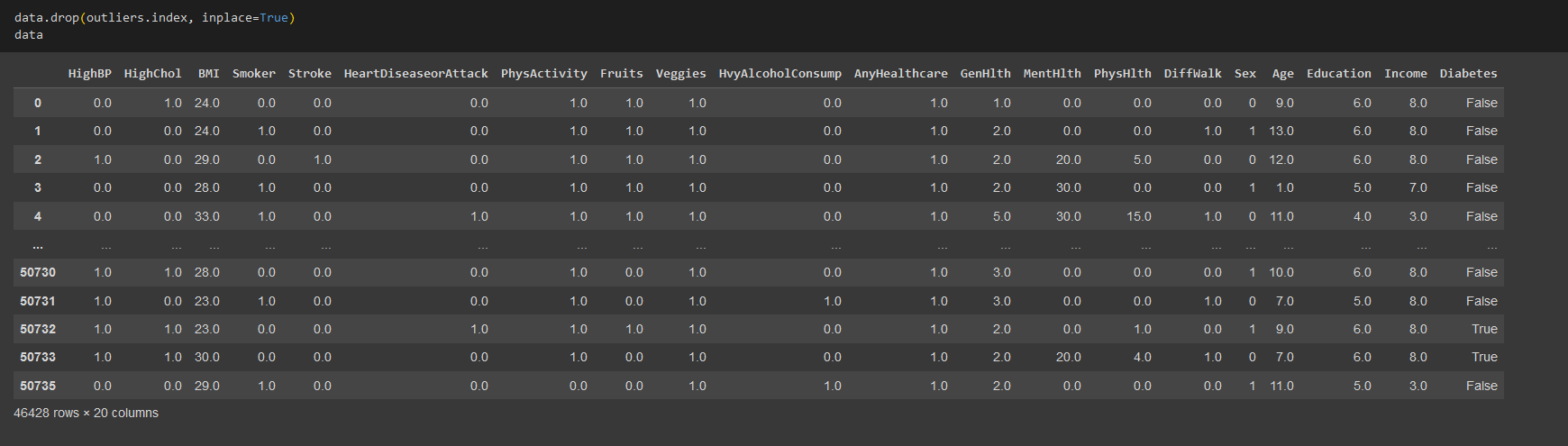
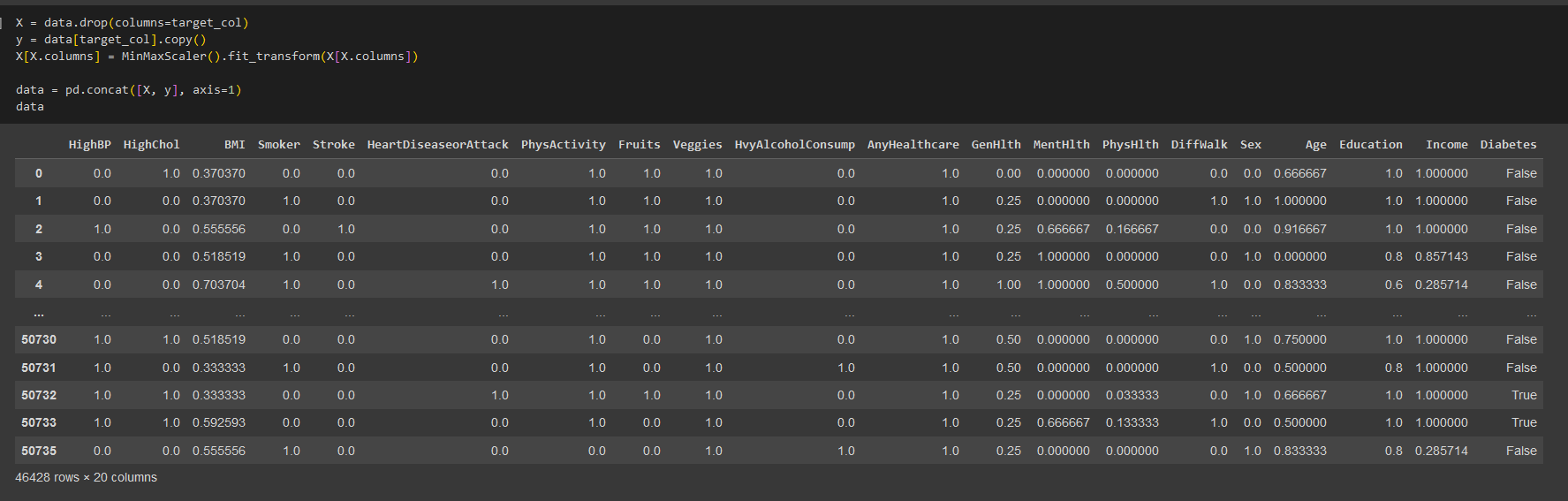
Berikut adalah tahapan modifikasi dataset untuk menangani isu-isu yang ditemukan pada poin 1-5.

* Tampilan dataset sebelum modifikasi (50736 baris).



* Encoding label pada kolom “Sex”.



* Penghapusan baris duplikat; jumlah baris tersisa adalah 48407.
* Identifikasi index baris outlier pada kolom “BMI”
* Pembuangan baris outlier; jumlah baris tersisa adalah 46428
* Normalisasi skala pada setiap kolom atribut. 

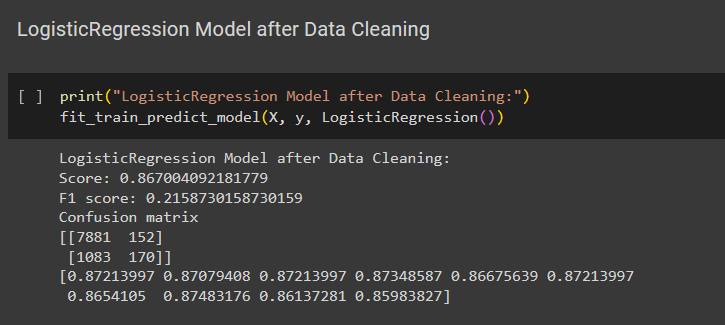
# DESAIN EKSPERIMEN

Strategi eksperimen penyusunan model meliputi pencarian grid search untuk mendapatkan nilai hyperparameter yang optimal. Strategi eksperimen juga meliputi pengujian model lain, misal RandomForest.

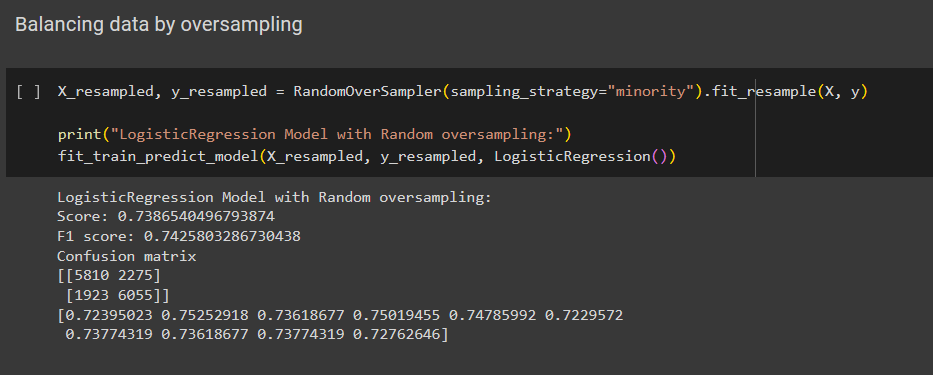
# HASIL EKSPERIMEN

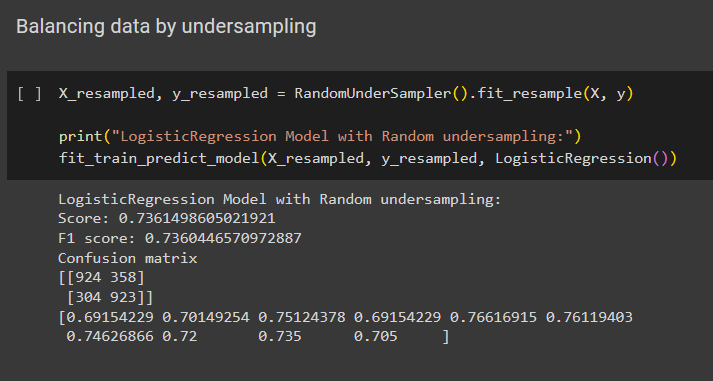
Berikut adalah hasil eksperimen menggunakan model-model yang disediakan oleh library Python. (Catatan: eksperimen dilakukan setelah proses penanganan isu data).

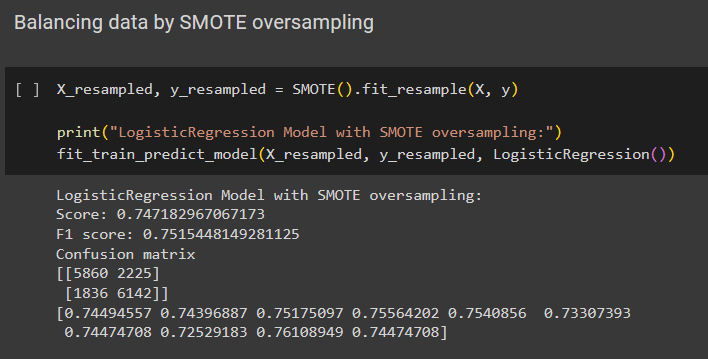
* Eksperimen menggunakan Logistic Regression



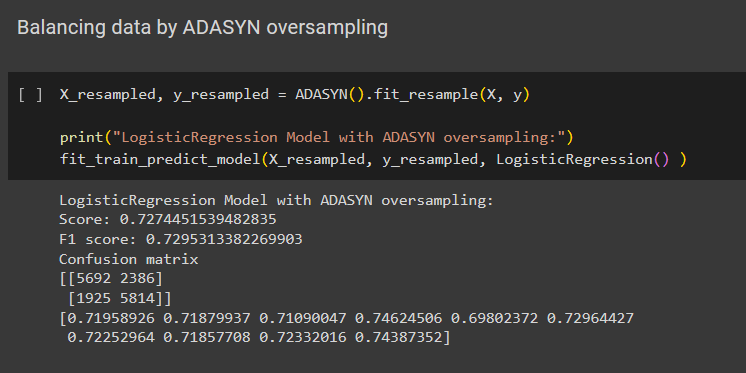
* Eksperimen oversampling menggunakan random oversampler



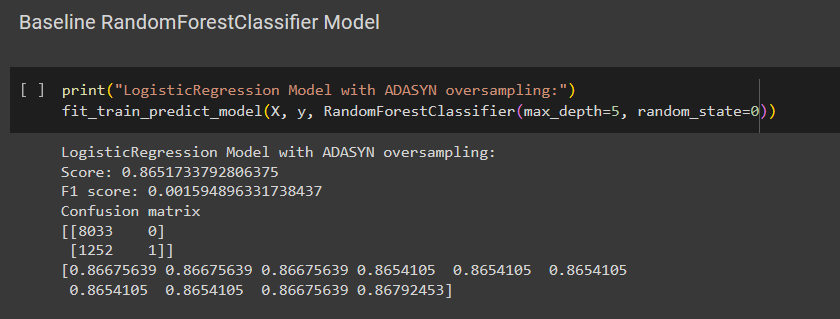
* Eksperimen undersampling menggunakan random undersampler
* Eksperimen oversampling menggunakan SMOTE



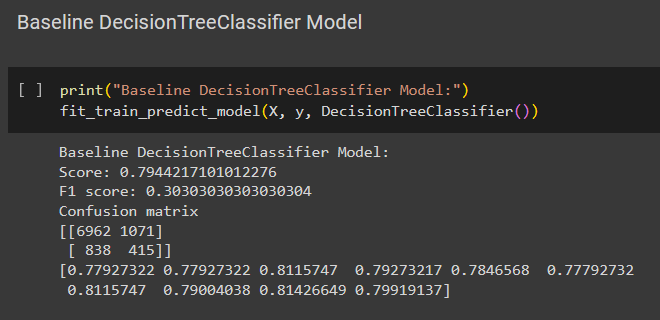
* Eksperimen oversampling menggunakan ADASYN



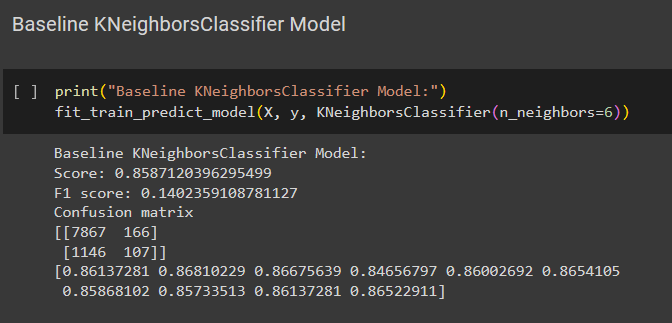
* Eksperimen model random forest classifier



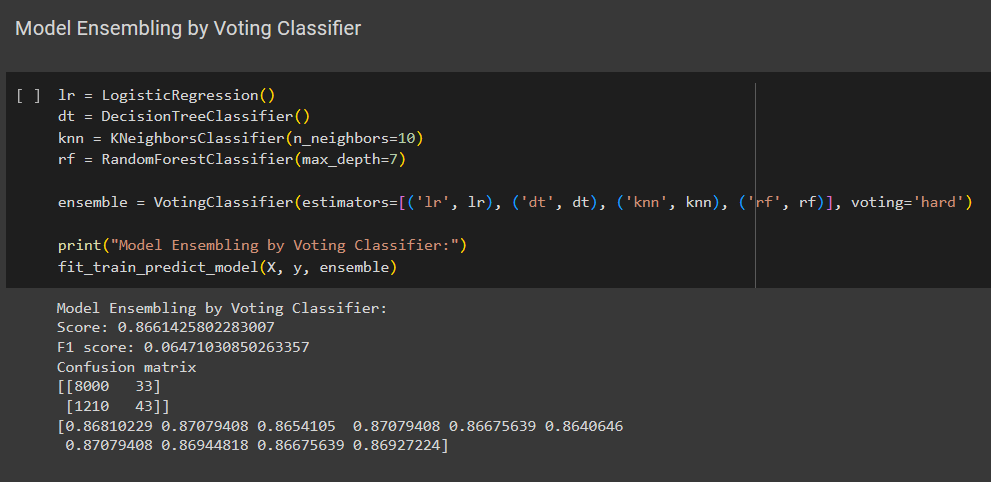
* Eksperimen model decision tree classifier



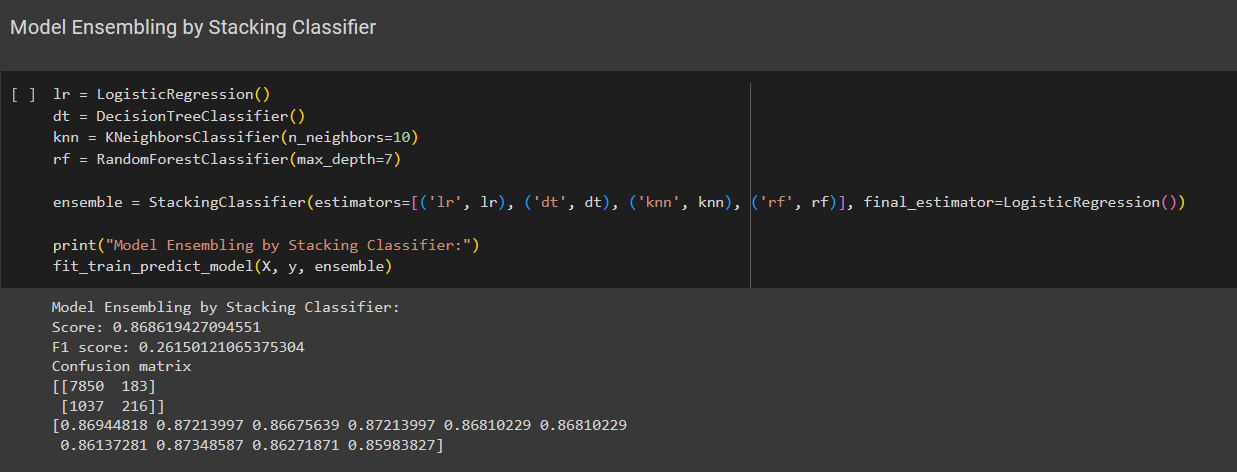
* Eksperimen model KNN



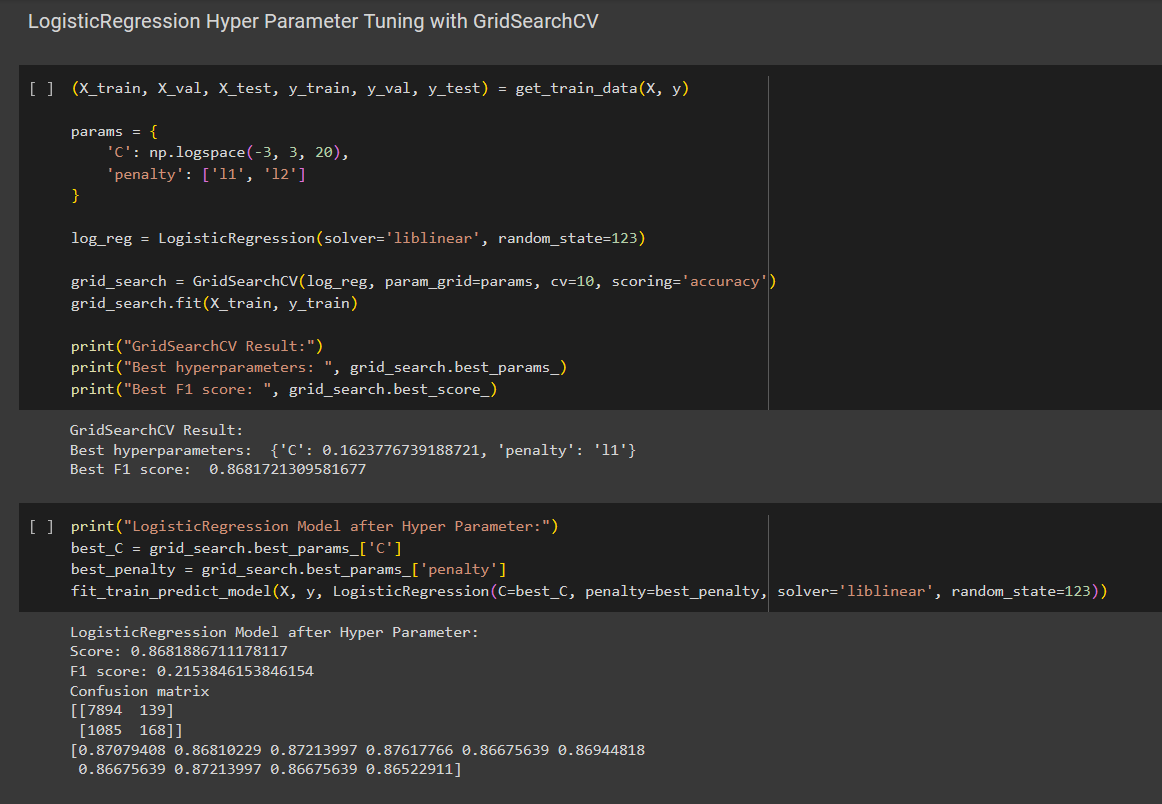
* Eksperimen model ensembling menggunakan voting classifier



* Eksperimen model ensembling menggunakan stacking classifier



* Eksperimen menggunakan grid search untuk hyperparameter tuning pada model logistic regression



# ANALISIS & KESIMPULAN DARI HASIL EKSPERIMEN

Berdasarkan hasil eksperimen menggunakan model LogisticRegression sebagai acuan dasar perbandingan nilai teknik-teknik dan model-model lain, kami menemukan bahwa model dengan skor tertinggi adalah model ensembler yang menggabungkan 4 buah model yaitu logistic regression, decision tree, k-nearest neighbors, dan random forest sebagai estimators. Penggabungan keempat estimator tersebut yang menghasilkan hasil terbaik adalah metode stacking classifier. Selain itu, percobaan penanganan imbalanced dataset dengan teknik oversampling dan undersampling menghasilkan nilai skor yang lebih rendah, tetapi memiliki F1-score yang lebih besar, sehingga model memiliki nilai precision dan recall yang besar.

Kesimpulan dari eksperimen ini adalah penyusunan model klasifikasi yang baik dapat dilakukan dengan bereksperimen dengan dataset, mengubah jenis model yang digunakan, menguji hyperparameter untuk setiap model, maupun menggabungkan beberapa model sekaligus menjadi satu model klasifikasi. Untuk membandingkan performa setiap model dapat dilakukan melalui analisis akurasi terhadap dataset pengujian, nilai F1-score, confusion matrix, dan skor validasi dari pengujian dataset validasi.

# 

# PEMBAGIAN TUGAS

| NIM | Nama | Tugas |
| --- | --- | --- |
| 13521074 | Eugene Yap Jin Quan | data cleaning (missing value, outlier), encoding, design of experiment |
| 13521173 | Dewana Gustavus Haraka Otang | data cleaning (imbalanced data, duplicate value), eksperimen model |