Deteksi Penipuan Menggunakan Random Forest

Dokumen ini merupakan versi bahasa indonesia dari proyek yang telah dilakukan pada:
https://www.kaggle.com/code/prakhosha/fraud-detection-using-random-forest

Pengantar

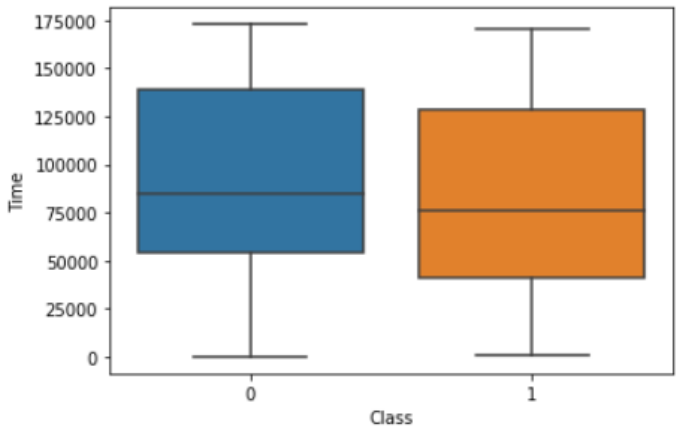
Pada projek ini, telah dilakukan deteksi penipuan menggunakan model pembelajaran mesin random forest pada library Sklearn. Data yang akan digunakan untuk projek ini dapat di akses pada:
https://www.kaggle.com/datasets/whenamancodes/fraud-detection

Data tersebut dari dari 31 kolom dengan satu di antara semua kolom tersebut merupakan kolom kelas. Satu kolom lainnya merupakan waktu pengguna untuk melakukan input data. Satu kolom lagi merupakan kolom jumlah uang. Sisa kolom yang tersisa merupakan keluaran dari algoritma PCA. Sayangnya karena faktor konfidensial, penyedia data tidak dapat membagikan informasi data masukan algoritma PCA.

Hasil

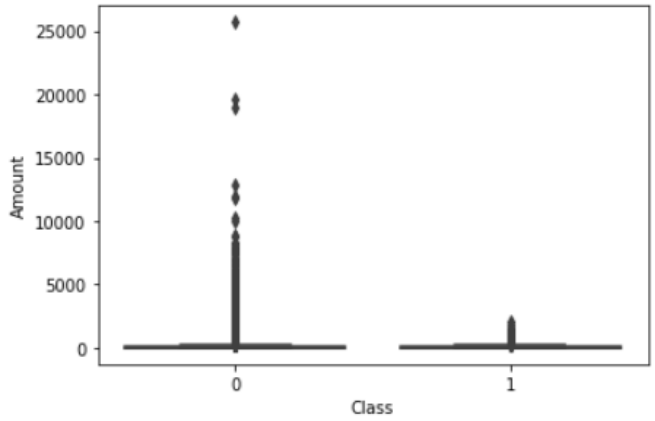
Hal pertama yang dilakukan adalah melakukan pembersihan data dari adanya duplikasi data, data yang hilang, dan sebagainya.

Selanjutnya dilakukan analisis boxplot variabel waktu pengguna untuk melakukan input data untuk mengetahui apakah variabel tersebut dapat digunakan dalam membentuk model pembelajaran mesin.



Secara visual, terlihat tidak ada perbedaan yang jelas untuk waktu input dari data penipuan dan data yang bukan merupakan penipuan. Oleh karena itu, variabel waktu input pengguna tidak akan dimasukan dalam model pembelajaran mesin.

Untuk variabel jumlah uang sendiri terlihat ada perbedaan yang jelas untuk kedua kelas seperti ada pada gambar di bawah. Oleh karena itu, variabel jumlah uang akan dimasukan ke dalam model pembelajaran mesin.



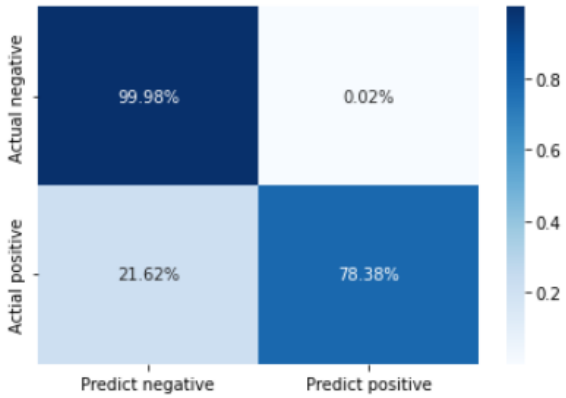
Data yang tersisa berjumlah 492 untuk data yang terindikasi adanya penipuan dan 284315 untuk data yang tidak terindikasi adanya penipuan. Karena jumlah data tidak seimbang, dibutuhkan pembobotan khusus dalam pembuatan model pembelajaran mesin.

Sebelum masuk ke pembentukan model, data yang tersedia dibagi menjadi dua set yaitu data untuk pelatihan model pembelajaran mesin dan model.

Selain itu, dilakukan standarisasi variabel dikarenakan variabel jumlah uang masih dalam dimensi yang tidak sama dengan fitur lainnya.

Untuk menemukan parameter yang akan digunakan dalam model pembelajaran mesin, digunakan teknik *grid search*. Setelah *grid search* dilakukan, ditemukan kualifikasi parameter yang terbaik untuk digunakan dalam model yaitu: 8 kedalaman pohon, entropi sebagai kriteria pembagi pohon, dan penggunaan bobot sub-sampel untuk menanggulangi ketidakseimbangan data.

Hasil pelatihan model menunjukkan F1 score sejumlah 0.81. Hasil confusion matriks dari prediksi data tes tes dapat dilihat pada gambar di bawah.



Terlihat bahwa model berhasil memprediksi 99.98% data yang bukan penipuan dan 78.38% data penipuan. Hasil ini merupakan hasil yang baik karena sebisa mungkin tidak ada data yang bukan penipuan tapi diklasifikasikan oleh model sebagai data penipuan.

Kesimpulan

Telah dibentuk model deteksi penipuan menggunakan algoritma random forest dengan performa model mencapai 0.81 F1 score dan model berhasil memprediksi 99.98% data yang bukan penipuan dan 78.38% data penipuan.