Prediksi Penangkapan Berdasarkan Data pada Chicago sejak 2001 Hingga 2018

Aditya Arya Hendrady, Arfi Renaldi, Harry Akbar Ali Munir, Rifki Adrian
Fakultas Ilmu Komputer
Universitas Indonesia

Abstrak

Hukum di Indonesia terbilang masih belum terlaksanakan dengan baik, banyak sekali jenis-jenis kejahatan dimanapun dan kapanpun yang terjadi di Indonesia. Bersalah atau tidaknya terduga kejahatan dapat ditentukan dari berbagai data-data yang diketahui dari suatu kejahatan yang telah terjadi. Penelitian ini ditujukan untuk mengembangkan sistem otomatis yang bisa memprediksi bersalah atau tidaknya pelaku suatu tindakan kriminal. Sistem ini menggunakan data-data yang ada yang akan menjadi penentu hasil akhir prediksi. Data yang digunakan adalah seluruh rekaman data dari Chicago yang peneliti nilai atribut yang tersedia cukup lengkap dan sangat menggambarkan kejahatan yang terjadi, sehingga dapat menjadi contoh yang baik untuk diterapkan di Indonesia. Sistem otomatis tersebut menggunakan kecerdasan buatan yang menggunakan beberapa algoritma klasifikasi yang akan membuat model klasifikasi, di antaranya adalah *Decision Tree* dan *Random Forest Classification*. Hasil prediksi yang menggunakan model klasifikasi tersebut adalah apakah suatu pelaku kejahatan dapat dinyatakan bersalah atau tidak.

Keyword: chicago crimes, klasifikasi, decision tree, data science, random forest

I. Pendahuluan

A. Latar Belakang

Di Indonesia, hukum penangkapan masih terbilang tidak terlaksanakan dengan baik dan tentu saja tidak adil. Banyak sekali kita lihat koruptor-koruptor yang tidak dijebloskan ke penjara disaat terdapat nenek renta tidak berpenghasilan yang ditangkap hanya karena sekadar mengambil buah dari pohon tetangga. Oleh karena itu dengan adanya model yang memprediksi penangkapan ini diharapkan dapat membantu menegakkan keadilan di Indonesia.

B. Rumusan Masalah

- Apakah algoritma klasifikasi yang berbeda akan menghasilkan akurasi prediksi yang berbeda pula?

C. Tujuan

Tujuan dari penelitian ini adalah untuk memprediksi apakah pelaku kejahatan dapat tertangkap dengan meninjau data kejahatan yang sudah terjadi sebelumnya. Dan juga untuk mengurangi ketidakadilan yang sering terjadi pada negara-negara *corrupt*.

D. Manfaat

Manfaat yang didapatkan pada penelitian ini adalah dapat memprediksi apakah pelaku kejahatan dapat tertangkap dengan meninjau data kejahatan yang sudah terjadi sebelumnya. Dan juga mengurangi ketidakadilan yang sering terjadi pada negara-negara *corrupt.*

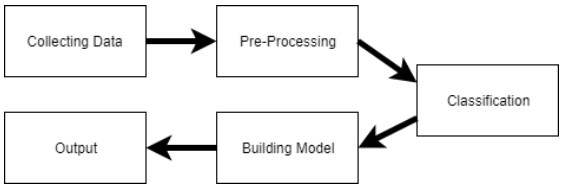
II. Data dan Metodologi

A. Data

Dataset kami peroleh berasal dari website resmi pemerintah Chicago berikut: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

Dataset tersebut berukuran sebesar 1.5GB dengan 22 fitur dan memiliki instance dari tahun 2001 hingga 2018. Fitur-fitur yang ada pada dataset menggambarkan tentang kejadian kejahatan yang pernah terjadi di Chicago, Amerika Serikat.

B. Metodologi



Gambar 1. Flow Chart Metodologi

Metodologi yang kami lakukan dalam penelitian ini adalah:

1. Collecting Data

Dataset yang diperoleh berdasarkan website resmi pemerintah Chicago, Amerika Serikat. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. Dataset tersebut berukuran sebesar 1.5GB. Fitur-fitur yang ada pada dataset menggambarkan tentang kejadian kejahatan.

2. Pre-processing

Pengolahan data dilakukan terlebih dahulu guna membersihkan data dari missing value, outlier, encoding feature menggunakan One Hot Encoding. Dan pada metode ini dilakukan sampling dan membandingkannya dengan data populasi, lalu mengecek null values, class balancing, mengecek outlier dan menghapus outlier, dan ekstraksi fitur dengan One Hot Encoding.

3. Building Model

Dilakukan beberapa metode klasifikasi menggunakan algortima machine learning, yaitu Decision Tree dan Random Forest.

4. Building Model

Pembuatan model dilakukan dengan menggunakan algoritma klasifikasi, salah satunya adalah Decision Tree. Pembuatan model ini sebelumnya dilakukan terlebih dahulu

5. Output

Setelah model dibangun, lalu model tersebut diuji dengan cara melakukan test pada model menggunakan data test. Lalu didapatkan nilai akurasi dari model tersebut.

C. Evaluasi

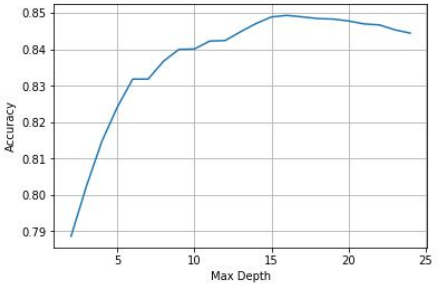
Metode evaluasi yang digunakan adalah dengan menghitung accuracy score menggunakan sklearn accuracy_score dan kemudian melihat confusion matrix dari output prediction dan output test.

```
Accuracy Max = 0.8493368779945628 at Max Depth = 16
True negative:  39930
True positive:  16618
False positive:  1491
False negative:  8540

Out[122]: array([[39930,  1491],
                 [ 8540, 16618]], dtype=int64)
```

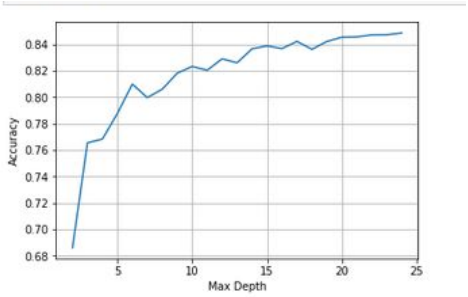
III. Hasil dan Analisis

Pada percobaan yang dilakukan menggunakan Decision Tree dan Random Forest, data diolah terlebih dahulu dengan melakukan metodologi yang telah dijelaskan sebelumnya. Tools yang digunakan salah satunya menggunakan library Pandas dan sklearn. Data dipartisi menjadi 30% data test dan 70% data training. Lalu dilakukan feature selection yang sama untuk semua algoritma. Berdasarkan hasil percobaan menggunakan Decision Tree, didapatkan hasil akurasi dengan membandingkan parameter max depth yang semakin besar, bisa dilihat pada grafik di bawah ini:



Gambar 2. Decision Tree

Lalu dilakukan percobaan menggunakan algoritma Random Forest dengan membandingkan parameter yang sama dan didapatkan hasil berikut:



Gambar 3. Random Forest

Berdasarkan grafik Decision Tree, max depth yang besar tidak menjamin akurasi yang meningkat, karena semakin besar max depth, maka akan banyak fitur yang akan digunakan dan ada beberapa fitur yang sebenarnya tidak perlu digunakan, sehingga semakin besar max_depth akan memungkinkan terjadi overfitting dan membuat akurasi menurun seperti pada gambar diagram 2.

Untuk Random Forest, diagram terlihat selalu naik seiring dengan membersarnya max_depth yang digunakan. Hal ini mengartikan bahwa Random Forest bisa mengatasi masalah yang ada pada Decision Tree, yaitu overfitting, karena pemilihan fitur yang dilakukan Random Forest dilakukan secara acak pada setiap tree yang dibuat.

Algoritma	Akurasi	Max Depth
Decision Tree	84.989%	16
Random Forest	84.86%	24

Tetapi, untuk mencapai akurasi maksimum pada Random Forest diperlukan max_depth yang lebih besar dibandingkan pada Decision Tree. Karena, pada Random Forest, pemilihan attribute/variable terbaik dilakukan pada variable yang dipilih secara random, sedangkan pada Decision Tree, pemilihan attribute/variable terbaik dilakukan sesuai dengan keadaan attribute terbaik saat itu.

IV. Kesimpulan dan Saran

Berdasarkan hasil penelitian dan analisis, kami menyimpulkan bahwa akurasi dari menggunakan Decision Tree tidak berbeda jauh dengan menggunakan Random Forest. Hanya saja, pada Decision Tree didapatkan hasil lebih cepat yaitu pada kedalaman tree 16, sedangkan pada Random Forest diperlukan kedalaman hingga 24.

Referensi

Crimes - 2001 to present | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Marsland, S. (2015). *Machine Learning: An algorithmic perspective*. Boca Raton, FL: CRC Press.