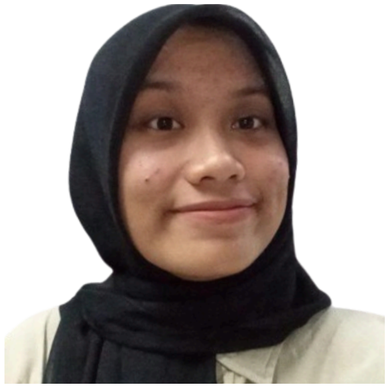
Implementasi Pyspark dengan Clustering K-Means dan Analisis Time Series Gradient Boosted Tree: Kasus Kecelakaan Kendaraan Bermotor United States

Kelompok 11 - Analisis Big Data

Our Super Team



Alyka Nazwa
121450016

Rizki Adrian Bennovry
121450073

Patricia Gaby R. T
121450099

Christian Arvianus B. N
121450112

Alber Analafean
121450146

Dataset

Motor Vehicle Collisions - Crashes

https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

Hasil dan Pembahasan

Tabel ini mengidentifikasi sektor-sektor rawan kecelakaan dengan frekuensi kecelakaan kendaraan paling sering, sehingga memungkinkan dilakukannya intervensi yang ditargetkan dan alokasi sumber daya untuk mengatasi masalah kecelakaan. Dengan membandingkan tingkat kecelakaan antar kota, pola dan faktor-faktor potensial yang mendasarinya dapat diidentifikasi sehingga dapat memandu pengambilan kebijaka, seperti kebijakan terkait keselamatan jalan raya, pembuatan peraturan lalu lintas yang lebih ketat, peningkatan infrastruktur, atau kampanye kesadaran masyarakat di kota-kota yang diidentifikasi.

Hotspot Analysis

Number of accidents by borough:

	BOROUGH	count
2	BROOKLYN	458,045
4	QUEENS	385,946
3	MANHATTAN	321,922
1	BRONX	212,906
5	STATEN ISLAND	60,439

CONTRIBUTING FACTOR VEHICLE 1	VEHICLE TYPE CODE 1	count
Unspecified	PASSENGER VEHICLE	226789
Unspecified	Sedan	146364
Driver Inattentio...	Sedan	140992
Driver Inattentio...	Station Wagon/Spo...	115160
Unspecified	Station Wagon/Spo...	103423

only showing top 5 rows

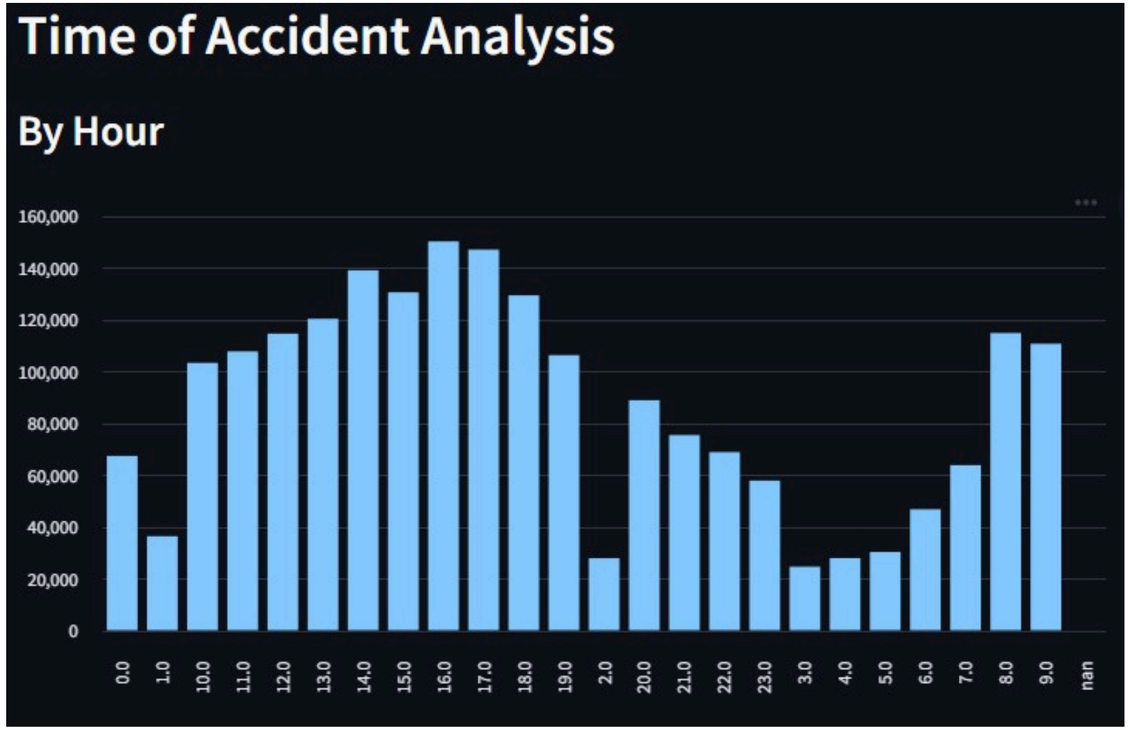
Tabel tersebut menampilkan informasi mengenai factor penyebab Utama dengan kendaraan penyebab Utama yang diurutkan dari yang paling sering dan ditampilkan 5 teratas. dapat diinterpretasikan bahwa jenis kendaraan passenger vehicle dan sedan dengan factor unspecified paling sering kemudian disusul oleh sedan dan station wagon dengan factor lalai dalam berkendara sehingga dapat dijadikan sebagai acuan dalam pembuatan peraturan

Terlihat pada table yang diberikan menunjukan jumlah kecelakaan terbanyak berdasarkan wilayah atau sektor dimana wilayah Brooklyn menempatkan posisi pertama dimana angka kecelakaan mencapai 458,045

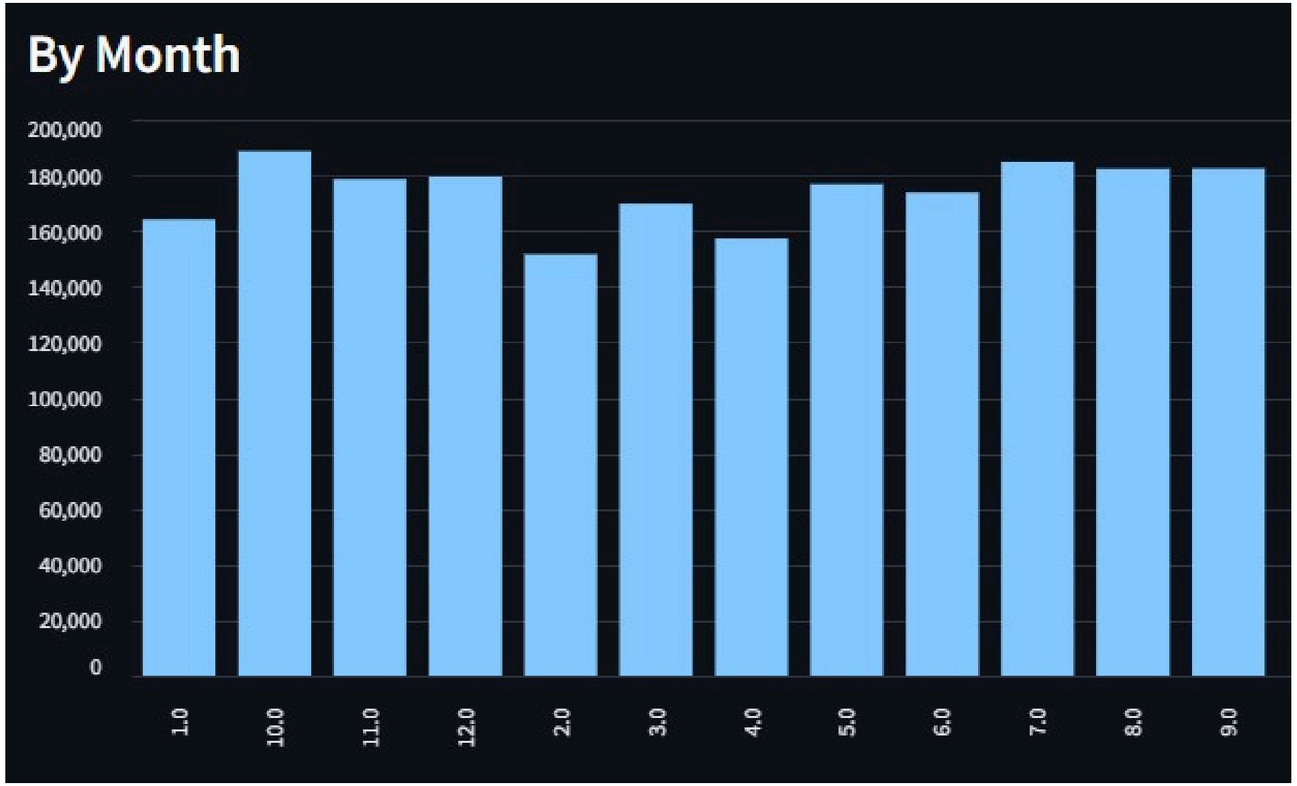
Causes of Accidents

Number of accidents by contributing factor:

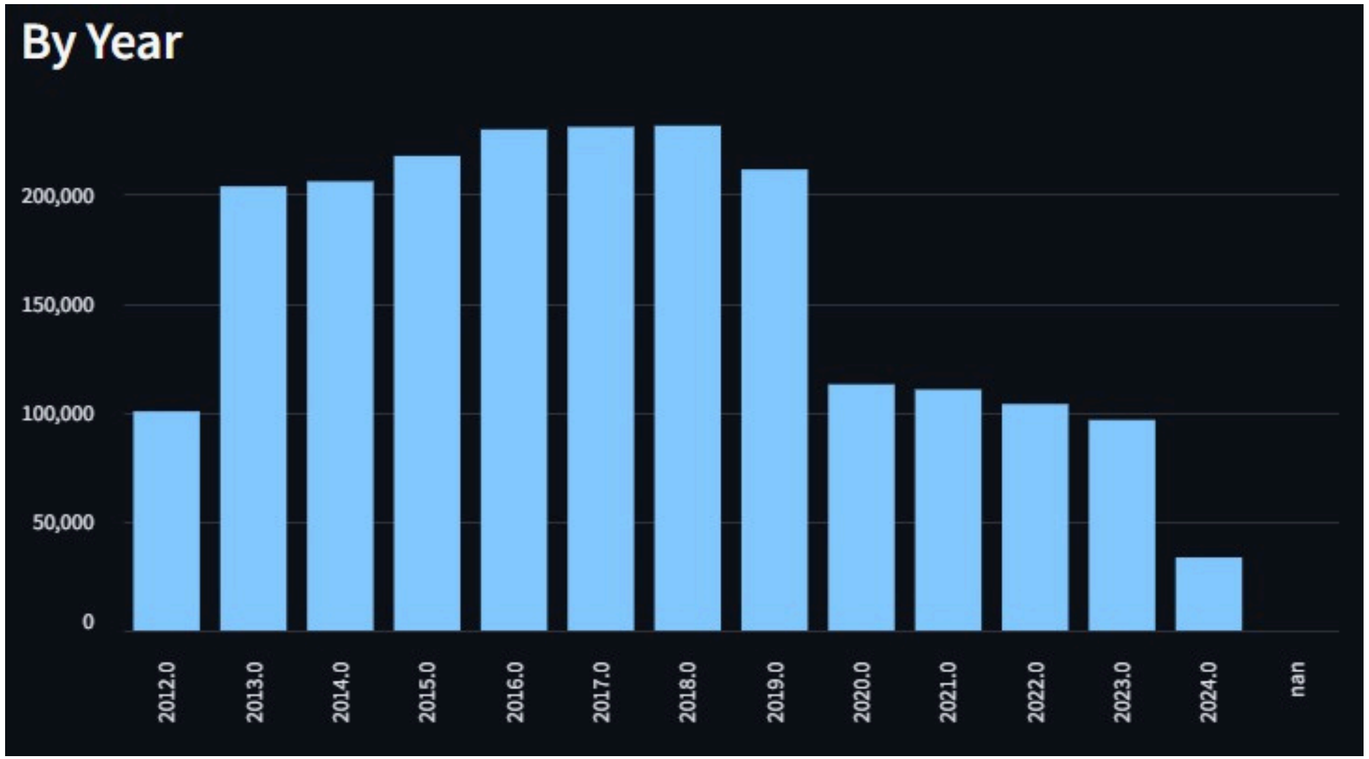
	CONTRIBUTING FACTOR VEHICLE 1	count
56	Unspecified	710,122
11	Driver Inattention/Distraction	418,621
18	Failure to Yield Right-of-Way	124,251
21	Following Too Closely	111,858
6	Backing Unsafely	77,146
32	Other Vehicular	64,841
37	Passing or Lane Usage Improper	58,421
36	Passing Too Closely	52,113
53	Turning Improperly	51,580
19	Fatigued/Drowsy	47,418



Grafik ini menunjukkan jam-jam kecelakaan paling sering terjadi yang dapat digunakan untuk mengoptimalkan manajemen lalu lintas dengan menyesuaikan waktu sinyal lalu lintas, mengerahkan polisi lalu lintas tambahan pada jam sibuk, atau menerapkan penutupan jalan sementara jika diperlukan untuk mengurangi kemacetan dan potensi kecelakaan dan alokasi sumber daya tim tanggap darurat dan personel medis secara lebih efektif untuk memastikan bantuan segera selama periode puncak kecelakaan.

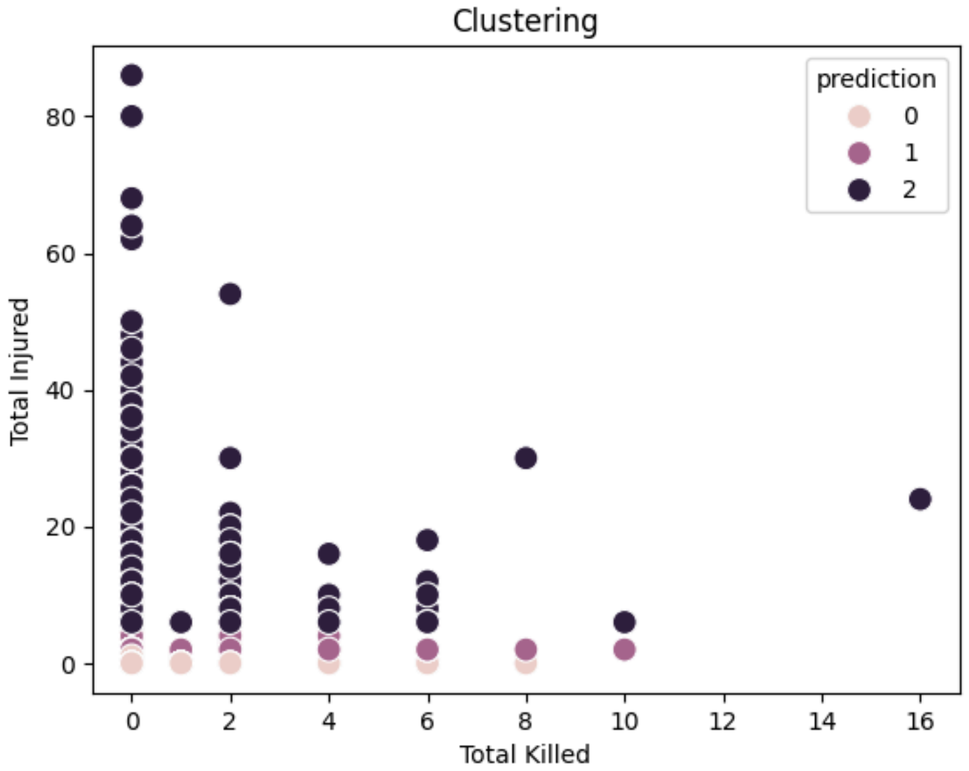


Terlihat pada table yang diberikan menunjukan jumlah kecelakaan terbanyak berdasarkan wilayah atau sektor dimana wilayah Brooklyn menempatkan posisi pertama dimana angka kecelakaan mencapai 458,045 terlihat pada Gambar (2)



Grafik ini memungkinkan untuk mengamati tren kecelakaan kendaraan secara keseluruhan selama bertahun-tahun. Dapat digunakan untuk menilai efektivitas langkah-langkah keselamatan kemudian evaluasi tentang peraturan yang lebih ketat atau infrastruktur jalan yang lebih baik, dan Identifikasi potensi penyebab fluktuasi dari waktu ke waktu.

hasil klustering yang dilakukan pada data kecelakaan berdasarkan jumlah korban terluka (Total Injured) dan jumlah korban tewas (Total Killed) dengan Tiga kluster yang berbeda diidentifikasi dengan warna yang berbeda:



Kluster 0 (Warna Terang/Pink Muda):

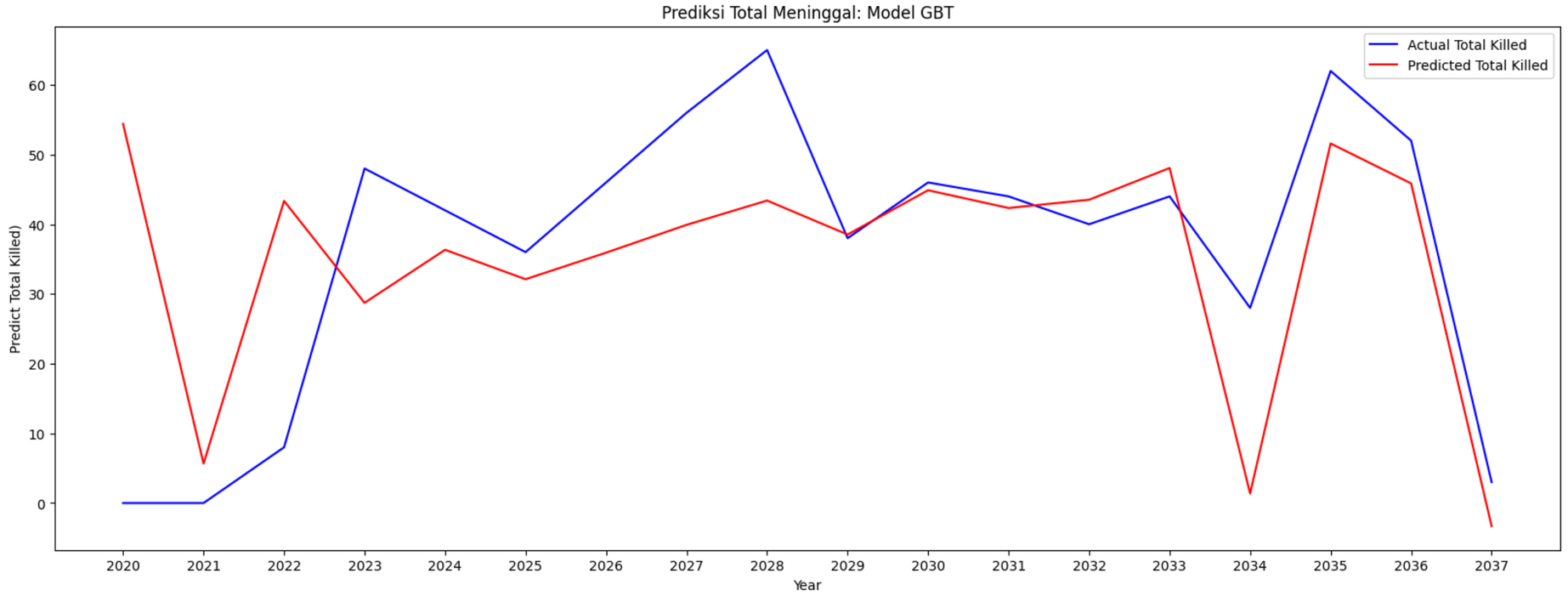
Kluster ini merepresentasikan kecelakaan dengan dampak ringan. Dengan karakteristik didonimasi data yang memiliki jumlah korban tewas yang rendah (0-2) dan korban terluka yang rendah (0-4). Ini menunjukkan bahwa kecelakaan dalam kelompok ini cenderung tidak terlalu parah.

Kluster 2 (Warna Gelap/Ungu Gelap):

Interpretasi: Kluster ini merepresentasikan kecelakaan dengan dampak berat. Dengan karakteristik data dalam kluster ini memiliki jumlah korban tewas yang lebih tinggi (hingga 16) dan korban terluka yang lebih tinggi (hingga 80). Kecelakaan dalam kelompok ini adalah yang paling parah, dengan angka korban tewas dan terluka yang signifikan.

Kluster 1 (Warna Ungu Muda):

Kluster ini merepresentasikan kecelakaan dengan dampak sedang. Dengan karakteristik data dalam kluster ini tersebar lebih luas dibandingkan kluster 0, dengan jumlah korban tewas berkisar antara 0 hingga sekitar 8 dan jumlah korban terluka hingga sekitar 20. Ini menunjukkan bahwa kecelakaan dalam kelompok ini memiliki variasi yang lebih besar dalam hal tingkat keparahan, tetapi umumnya lebih parah daripada kluster 0.


Prediksi Total Meninggal: Model GBT
Actual Total Killed
Predicted Total Killed
Predict Total Killed)
Year
0
10
20
30
40
50
60
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037

Kesimpulan

Penelitian kali ini peneliti mengambil metode *K-Means Clustering* dan *Gradient Boost Tree*, dengan tujuan utama dari penelitian ini adalah untuk mengidentifikasi pola-pola tersembunyi dalam data kecelakaan dan memprediksi tren masa depan guna meningkatkan kebijakan keselamatan lalu lintas. Algoritma K-Means digunakan untuk mengelompokkan data kecelakaan berdasarkan karakteristik tertentu, sedangkan Gradient Boosted Tree diterapkan untuk melakukan analisis prediktif terhadap data kecelakaan. Berdasarkan penelitian yang telah dilakukan, didapati bahwa metode *K-Means Clustering (Clustering)* dengan Gradient Boosted Tree *(Ensemble Learning)* dapat dikombinasikan sehingga menghasilkan model yang kuat.

Thank You