Original Article

Received 00th January 20xx
Accepted 00th Febuary 20xx
Published 00th March 20xx

Open Access

DOI: 10.35472/x0xx0000

Implementasi Pyspark dengan Clustering K-Means dan Prediksi Gradient Boosted Tree: Kasus Kecelakaan Kendaraan Bermotor di United States

Christian Arvianus Natanael Biran[a], Alber Analafean[b], Patricia Gaby Rahmawati Tamba[c], Rizki Adrian Bennovry[d], Alayka Nazwa[e]

[a] Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung Selatan, Lampung

** Corresponding E-mail: christian.1214500112@student.itera.ac.id*

Abstract: In this study, we implemented Pyspark to perform clustering using the K-Means algorithm and analyze time series data with the Gradient Boosted Tree method in the case of motor vehicle accidents in the United States. The main objective of this research is to identify hidden patterns in accident data and predict future trends to improve traffic safety policies. The K-Means algorithm is used to cluster accident data based on certain characteristics, while the Gradient Boosted Tree is applied to perform predictive analysis on the accident time series data. The research results show that the combination of these two methods can provide deep insights and accurate predictions regarding motor vehicle accidents in the United States.

Keywords: *Pyspark, K-Means Clustering, Time Series Analysis Gradient Boosted Tree, Motor Vehicle Accidents*

Abstrak: Dalam penelitian ini, dilakukan pengimplementasian pengolahan big data menggunakan Pyspark untuk melakukan clustering menggunakan algoritma K-Means dan prediksi data time series dengan metode Gradient Boosted Tree pada kasus kecelakaan kendaraan bermotor di United States. 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 time series kecelakaan. Hasil penelitian menunjukkan bahwa kombinasi kedua metode ini dapat memberikan wawasan mendalam dan prediksi yang akurat mengenai kecelakaan kendaraan bermotor di United States.

Kata Kunci: *Pyspark, Clustering K-Means, Analisis Time Series, Gradient Boosted Tree, Kecelakaan Kendaraan Bermotor.*

Pendahuluan

Kecelakaan lalu lintas tetap menjadi global yang belum terpecahkan. Menurut data dari WHO, setiap detiknya setidaknya satu orang yang meninggal akibat kecelakaan, dan jumlah kasus kecelakaan terus meningkat sebesar 100 ribu setiap tiga tahun. Berdasarkan laporan terbaru pada tahun 2013, ada 1,25 juta orang meninggal karena kecelakaan lalu lintas di seluruh dunia [1]. Dari permasalahan tersebut didapatkan alternatif untuk dapat menurunkan angka kecelakaan di lalu lintas, seperti pembatasan kecepatan maksimal, larangan mengemudi dengan keadaan mabuk. Data WHO mencatat bahwa

negara-negara dengan tingkat pendapatan perkapita tinggi seperti Amerika, jumlah kematian akibat kecelakaan lalu lintas menempati peringkat ke-14 dengan kematian rata-rata 15,0 per 100.000 penduduk. Negara-negara dengan tingkat pendapatan perkapita rendah cenderung memiliki prevalensi yang lebih tinggi yaitu menempati urutan ke-10 penyebab kematian[2]. Terdapat beberapa faktor yaitu : 1) Kecepatan saat tabrakan, faktor yang berpengaruh terhadap beratnya trauma adalah kecepatan gabungan antara sepeda motor dan lawan tabrak. jika

kecepatan gabungan >100 km/jam maka *Risk Prevalen* (RP) adalah 3,3 kali lebih besar untuk mengalami cedera berat dibandingkan ≤ 100 km/jam [3], 2) Penggunaan helm berstandar, hasil penelitian yang dilakukan menunjukkan bahwa beratnya trauma /cedera akibat kecelakaan lalu lintas pada pengemudi sepeda motor disebabkan oleh penggunaan helm substandar (non standar) dan tidak ditali (unchain strap) dengan prevalensi risiko sebesar 4,833 untuk helm substandar dan 2,143 untuk tidak ditali, 3) Konsumsi alkohol, kadar alkohol saat mengemudi menurunkan tingkat kewaspadaan dan keseimbangan mengemudi sepeda motor. Kombinasi antara tidak menggunakan helm dan dalam keadaan minum beralkohol memiliki kecenderungan mengalami traumatis yang lebih parah [4], 4) Tipe kecelakaan, tipikal kecelakaan yang terjadi mempengaruhi tingkat keparahan pengendara, tipe kecelakaan tubrukan (berhadapan) di salah satu jalan arteri di Bali mencapai 89,1% dan 44% dari tubrukan tersebut mengalami cedera fatar [5], 5) Perilaku agresif, keparahan cedera kepala dipengaruhi oleh perilaku agresif pengendara. Diantara perilaku agresif adalah merancang ekstrim sepeda motor dan fakta membuktikan bahwa 78% kecelakaan terjadi di Amerika Serikat dikarenakan hal tersebut [6].

Terdapat beberapa referensi penelitian sebelumnya yang telah menerapkan metode *K-Means* pada kasus kecelakaan kendaraan bermotor, salah satu peneliti tersebut adalah S. Ahmed, M.S dan M.A. Habib dimana mereka melakukan penelitian dengan judul *"An Analysis of Motor Vehicle Accident in the United States Using K-Means Clustering"*, dimana dalam penelitian ini peneliti menggunakan *K-Means* untuk mengelompokkan kecelakaan kendaraan bermotor di *United States* berdasarkan faktor-faktor seperti lokasi, waktu, dan jenis kecelakaan. Hasil penelitian menunjukkan bahwa metode *K-Means* dapat digunakan untuk mengidentifikasi pola dan trend dalam data kecelakaan kendaraan bermotor, yang dapat membantu dalam pengembangan strategi pencegahan kecelakaan [7]. Peneliti selanjutnya H.Wang, Y.sun, dan X.Ma telah melakukan penelitian serupa dengan judul penelitian *"K-Means Clustering for identifying High-Risk Areas of Motor Vehicle Accident in the United States"*, dimana penelitian ini mengidentifikasi area berisiko tinggi kecelakaan kendaraan bermotor di *United States*. Hasil penelitian menunjukkan bahwa metode *K-Means* digunakan untuk mengidentifikasi area yang memiliki tingkat kecelakaan kendaraan bermotor yang lebih tinggi, yang dapat membantu dalam mengalokasikan sumber daya untuk pencegahan kecelakaan [8]. Peneliti J.Smith, M.Johanes, dan D. Williams juga pernah melakukan penelitian serupa dengan judul *"Using K-Means Clustering to Analyze Motor Vehicle Accident Data in the United States: A Case Study"*, di dalam penelitian mereka menganalisis data dimana hasil penelitian menunjukkan bahwa metode *K-Means* dapat digunakan untuk mengidentifikasi faktor - faktor yang berkontribusi terhadap kecelakaan kendaraan bermotor di *United States*, yang dapat membantu dalam pengembangan strategi pencegahan kecelakaan di *United States* [9].

Berdasarkan informasi dan penelitian yang telah dilakukan sebelumnya, terlihat bahwa kecelakaan lalu lintas masih menjadi permasalahan global yang serius. Berbagai upaya yang telah dilakukan untuk menurunkan angka kecelakaan, namun hasilnya belum optimal. Oleh karena itu, penelitian ini dilakukan bertujuan untuk melakukan *clustering* kecelakaan bermotor di *United States* dimana akan dikelompokkan menjadi tiga klaster yaitu, kluster kecelakaan berat mengelompokkan data kecelakaan yang mengakibatkan korban jiwa atau luka berat, kluster kecelakaan sedang mengelompokkan data kecelakaan yang yang mengakibatkan luka ringan, lalu yang ketiga kluster kecelakaan ringan mengelompokkan data kecelakaan yang tidak mengakibatkan korban jiwa atau luka. Lalu penelitian ini akan menghasilkan hasil *clustering* untuk mengetahui *cluster* mana yang paling sering terjadi di US. Hasil ini dapat membantu dalam menentukan strategi pencegahan kecelakaan yang lebih tepat. Lalu memvisualisasikan hasil *clustering* menggunakan teknik visualisasi data yang menarik. Hal ini dapat membantu dalam memahami pola dan sebaran data kecelakaan dengan lebih mudah.

Selain itu, pada penelitian ini menggunakan metode *Gradient Boosted Tree (GBT)* yang akan digunakan untuk melakukan analisis data *time series* dan analisis prediktif yang bertujuan memprediksi tren kecelakaan masa mendatang berdasarkan pola historis. Metode ini bekerja dengan membangun serangkaian model prediksi yang mampu menangkap kompleksitas dan pola *non-linear* dalam data *time series*. Dengan memahami prediksi tren ini, diharapkan dapat mengambil tindakan preventif yang lebih efektif untuk mengurangi kecelakaan di masa depan. Hasilnya dapat menjadi bahan evaluasi untuk kepolisian dalam menentukan strategi pencegahan kecelakaan. Penelitian ini akan menggunakan *PySpark*, sebuah *framework* pemrosesan data terdistribusi, untuk mengolah data kecelakaan yang berukuran besar. Hal ini dapat meningkatkan efisiensi dan kecepatan analisis data. Dengan mencapai tujuan-tujuan tersebut, diharapkan penelitian ini dapat memberikan kontribusi dalam upaya menurunkan angka kecelakaan lalu lintas di US.

Metode

1. Pemrosesan Awal Data

   Pemrosesan Awal Data adalah proses ketika data yang akan digunakan akan dilakukan manipulasi, pengubahan, dikodekan, dan lain-lain untuk dianalisis dalam suatu algoritma pembelajaran mesin. Pemrosesan awal data dapat membantu algoritma pembelajaran mesin menguraikannya dengan cepat, sehingga proses analisis prediktif dapat dilakukan

a. Pemilihan Kumpulan

Pemilihan data yang akurat sangat penting dalam analisis, khususnya dengan berbagai alasan seperti kualitas data, relevansi data, jenis data, volume data, distribusi data, multikolinearitas, skalabilitas, dan tujuan peneliti dalam analisis

Dalam penelitian ini, peneliti menggunakan data Kecelakaan Kendaraan Bermotor di US yang tervalidasi melalui *data.cityofnewyork.us* dan telah di*update* pada 17 Mei 2024. Data ini memuat rincian kejadian kecelakaan kendaraan bermotor yang dilaporkan polisi di *New York City. Dataset* tersebut terdiri dari 1.048.556 baris dengan 29 kolom diantaranya, yaitu:

1. *Crash Date* (Tanggal Kecelakaan)
2. *Crash Time* (Waktu Kecelakaan)
3. *Borough* (Wilayah)
4. *Zip Code* (Kode Pos)
5. *Latitude* (Garis Lintang)
6. *Longitude* (Garis Bujur)
7. *Location* (Lokasi)
8. *On Street Name* (Nama Jalan)
9. *Cross Street Name* (Nama Lintas Jalan)
10. *Off Street Name* (Nama Jalan Luar)
11. *Number of Persons Injured* (Jumlah Orang yang Cedera)
12. *Number of Persons Killed* (Jumlah Orang yang Meninggal)
13. *Number of Pedestrians Injured* (Jumlah Pejalan Cedera)
14. *Number of Pedestrians Killed* (Jumlah Pejalan Meninggal)
15. *Number of Cyclist Injured* (Jumlah Pengendara Sepeda yang Cedera)
16. *Number of Cyclist Killed* (Jumlah Pengendara Sepeda yang Meninggal)
17. *Number of Motorist Injured* (Jumlah Pengendara Motor yang Cedera)
18. *Number of Motorist Killed* (Jumlah Pengendara Motor yang Meninggal)
19. *Contributing Factor Vehicle 1* (Faktor Penyedia Kendaraan 1)
20. *Contributing Factor Vehicle 2* (Faktor Penyedia Kendaraan 2)
21. *Contributing Factor Vehicle 3* (Faktor Penyedia Kendaraan 3)
22. *Contributing Factor Vehicle 4* (Faktor Penyedia Kendaraan 4)
23. *Contributing Factor Vehicle 5* (Faktor Penyedia Kendaraan 5)
24. *Collision_ID*
25. *Vehicle Type Code 1* (Kode Jenis Kendaraan 1)
26. *Vehicle Type Code 2* (Kode Jenis Kendaraan 2)
27. *Vehicle Type Code 3* (Kode Jenis Kendaraan 3)
28. *Vehicle Type Code 4* (Kode Jenis Kendaraan 4)
29. *Vehicle Type Code 5* (Kode Jenis Kendaraan 5)

b. Pemilihan Pemodelan

Peneliti akan menggunakan metode *K-Means Clustering* dan *Gradient Boost Tree*. Metode *K-Means Clustering* umumnya digunakan pada studi kasus pengelompokkan kumpulan data. Sedangkan, *Gradient Boosted Tree* umumnya digunakan dalam membuat model pohon keputusan dengan dasar model *decision tree*, namun *Gradient Boosted Tree* dibangun berdasarkan beberapa model yang lemah. Selain berfungsi membangun model pohon keputusan, *Gradient Boosted Tree* bisa digunakan dalam analisis *time series* dalam memprediksi total meninggal pada kasus kecelakaan kendaraan bermotor.

Pada Metode *K-Means Clustering*, peneliti ingin mengelompokkan tingkat kecelakaan menjadi tiga tingkat (Berat, Sedang, Ringan). Pengelompokkan dilakukan berdasarkan kolom "*Total Killed"* dan "*Total Injured"*. Sedangkan, pada Metode *Gradient Boosted Search*, peneliti ingin memprediksi Total korban kecelakaan yang meninggal di tahun selanjutnya berdasarkan pengelompokkan menggunakan *K-Means Clustering* yang sudah dilakukan sebelumnya.

c. Langkah-langkah

Hal pertama yang dilakukan, yaitu dengan melakukan *data preprocessing. Data preprocessing* yang dilakukan yaitu menemukan nilai yang unik pada setiap masing-masing kolom data, kemudian menghitung frekuensi nilai berdasarkan kelompok *"Bourough"* dan mengurutkannya berdasarkan *count* dengan pengurutan data dari nilai

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

terbesar ke terkecil *(descending)*. Setelah itu, mengkonversikan format kolom *"CRASH DATE"* menjadi tipe data dengan format yang sesuai yaitu *"dd/MM/yyy".* Dengan hal ini, kita dapat menambahkan kolom *"YEAR"* (Tahun) dengan menggunakan fungsi year. Agar secara keseluruhan, peneliti dapat melihat Jumlah Pengendara yang Terluka dan Jumlah Pengendara yang Meninggal, maka peneliti melakukan *summing up* di setiap atribut yang mengandung *INJURED* dan *KILLED* sehingga dibentuklah kolom baru pada yaitu *'Total Injured'* dan *'Total Killed'.* Kolom *'Total Injured'* dan *'Total Killed'* diagregasikan dengan kolom pada data awal (secara keseluruhan). Maka, kolom yang setelah diagregasikan sekarang yaitu *'YEAR', 'COLLISION_ID', 'BOROUGH', 'Total Killed', 'Total Injured', 'CONTRIBUTING FACTOR VEHICLE 1'.* Langkah terakhir dalam *data preprocessing*, yaitu menangani hilang dalam kolom *'Total Killed'* dan *'Total Injured'* pada DataFrame 'df' dengan melakukan imputasi nilai yang *NULL* atau hilang menggunakan nilai yang dihitung (misal, mean dan median). Hal ini bertujuan untuk membantu dalam meningkatkan kualitas data dan memastikan data lengkap dan siap untuk pemodelan lebih lanjut.

Hal kedua, membuat *features assemble*, dimana *library* yang digunakan yaitu *VectorAssembler*. Dengan membuat objek *VectorAssembler* yang dimana input kolom nya yaitu *"Total KIlled"* dan *"Total Injured"* dengan nama kolom hasil vektor fitur gabungan dari kolom "*Total KIlled"* dan *"Total Injured"* yaitu kolom *"features"*. Kemudian, dilakukan transformasi pada DataFrame 'df' dengan menggunakan objek *VectorAssembler*.

Langkah ketiga, membuat model *K-Means*. model *K-Means* yang dibuat sebanyak tiga klaster, yaitu klaster Berat, Sedang, dan Ringan.Kemudian dilakukan prediksi *K-Means* dengan mentransformasikan model pada *"Total Killed"* dan *"Total Injured"*. Kemudian, hasil dari prediksi tersebut diubah formatnya ke dalam Pandas.

Kemudian, langkah terakhir yang dilakukan pada metode *K-Means Clustering* yaitu membuat visualisasi dengan menggunakan *matplotlib* dan *seaborn* untuk menunjukkan total korban kecelakaan per tahun. Plot ini dilengkapi dengan judul, label sumbu, dan legenda untuk mempermudah interpretasi dan pemahaman visualisasi.

Setelah diperoleh model *clustering* menggunakan model *K-Means,* selanjutnya kita melakukan prediksi total korban kecelakaan pada tahun berikutnya dengan menggunakan *Gradient Boosted Tree*. Hal pertama dilakukan dalam prediksi menggunakan *Gradient Boosted Tree*. Dilakukan tahapan *preprocessing*, mengubah format tipe data pada kolom "*CRASH DATE*" ke format "mm/dd/yyyy. Tahapan selanjutnya dilakukan eksplorasi data pada beberapa atribut. Pada eksplorasi data dilakukan pada atribut *"Borough", "Hour", "CONTRIBUTING FACTOR VEHICLE", "VEHIICLE TYPE CODE 1", "CONTRIBUTING FACTOR VEHICLE 1".* Eksplorasi data yang dilakukan untuk mengetahui jumlah dari beberapa atribut yang memiliki pengaruh terhadap kasus kecelakaan. Selanjutnya dipilih kolom yang akan digunakan untuk membuat model *Gradient Boosted Tree* yaitu *'CRASH DATE', 'YEAR', 'MONTH', 'HOUR', 'COLLISION_ID', 'BOROUGH', 'NUMBER OF PERSONS INJURED', 'NUMBER OF PERSONS KILLED', 'NUMBER OF PEDESTRIANS INJURED', 'NUMBER OF PEDESTRIANS KILLED', 'NUMBER OF CYCLIST INJURED', 'NUMBER OF CYCLIST KILLED', 'NUMBER OF MOTORIST INJURED', 'NUMBER OF MOTORIST KILLED', d*an *'CONTRIBUTING FACTOR VEHICLE 1'.*

Selanjutnya dilakukan agregasi kolom *'Total Injured' deng*an beberapa kolom terkait luka yaitu *'NUMBER OF PEDESTRIANS INJURED','NUMBER OF CYCLIST INJURED','NUMBER OF MOTORIST INJURED' d*an juga untuk kolom *'Total Killed' y*aitu dengan kolom *'NUMBER OF PERSONS KILLED', 'NUMBER OF PEDESTRIANS KILLED','NUMBER OF CYCLIST KILLED','NUMBER OF MOTORIST KILLED'*. Hasil agregasi dijadikan ke dalam *dataframe timeseriesDf*. *Data Frame timeseriesDf* ini berisi atribut hasil agregasi yaitu *Month, Year, sum(Total Injured), sum(Total Killed*) yang digunakan membuat prediksi dengan menggunakan *Gradient Boosted Tre*e. Selanjutnya dilakukan interpolasi pada data yang memiliki *missing values* agar model yang dibangun akurat. Selanjutnya split data menjadi train set dan test set mengunakaan *timeseriesDf*, pada train_set menggunakan data *timeseriesDf* yang difilter pada tahun kurang dari 2020 dan pada test_set menggunakan data *timeseriesDf* yang difilter pada tahun lebih dari 2020. Selanjutnya dilakukan vektor assembler menggunakan atribut *"sum(Total Killed)"* dan *"sum(Total Injured)"* yang diubah kedalam vektor dan dilakukan korelasi menggunakan *pearson* dan *spearman*.

Selanjutnya dilakukan *feature scalling* berfungsi dalam menormalisasikan data serta mengoptimalkan performa model. Dalam melakukan skalalisasi fitur dengan menggunakan fungsi *tanh estimator* dan juga *scale_transform*. Selanjutnya dilakukan *sliding window* yang berfungsi mempartisi data tersebut ke beberapa *window* dalam mempermudah mengolah data dengan berukuran besar ini dan efisiensi pengolahan data, dimana *window* diatur sebanyak 30 . Selanjutnya dilakukan merge nilai x dan y dengan memanfaatkan fitur RDD.zip yang memanfaatkan *Resilient Distributed Dataset(RDD)* yang berfungsi dalam

mempartisi silang *multiple node* di dalam *cluster.* Selanjutnya dilakukan vektorisasi *window* yang berfungsi dalam melakukan transformasi *sequence data* kedalam fitur tunggal yang menyesuaikan algoritma.

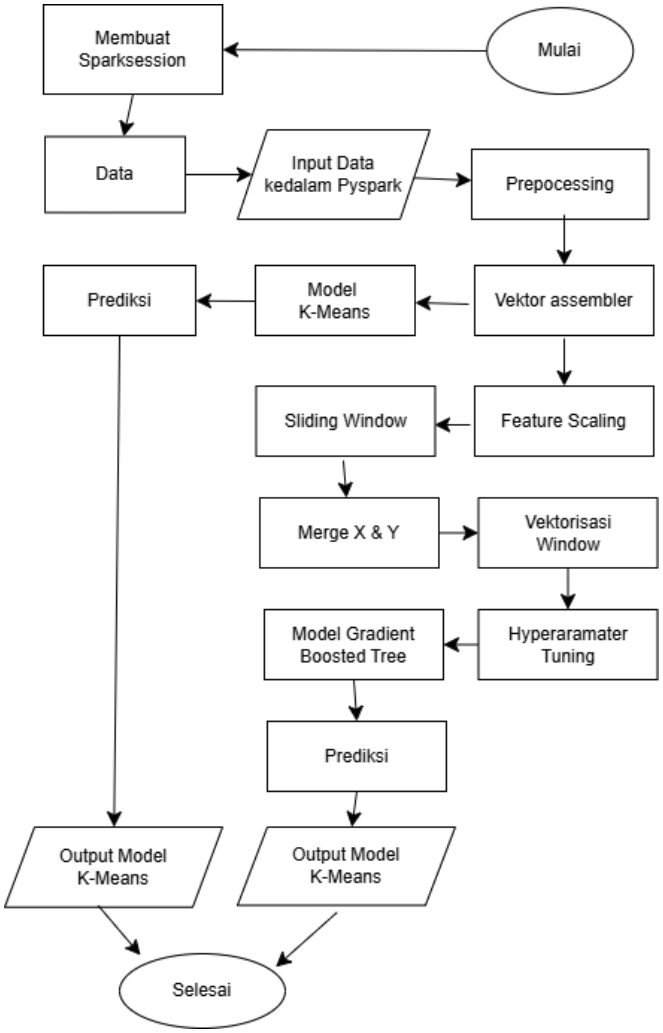
Selanjutnya dilakukan efisiensi algoritma agar model bisa dijalankan secara cepat dengan data yang berukuran besar tersebut dengan *hyperparameter tuning*. *Hyperparameter tuning* menggunakan yaitu *RegressionEvaluator* dan *CrossValidator.*

Setelah dilakukan proses *hyperparameter tuning* , dilakukan *Model Gradient Boosted Tree* dibangun dengan menggunakan data frame *timeseriesDF*, pada model yang dibangun menggunakan kolom *"sum(Total Killed)"* dalam memprediksi total korban kecelakaan pada tahun yang akan datang. Pada model yang dibangun dilakukan *window* Spesifikasi. Fungsi-fungsi ini beroperasi pada baris dalam *window* tertentu, mengekstraksi informasi dan melakukan penghitungan berdasarkan titik data di sekitarnya. *window* spesifikasi menentukan bingkai *window* yang tepat dimana operasi ini diterapkan. Selanjutnya dilakukan vector assembler dengan inputan kolom *"lag1"* dan *"lag2"* dengan output kolom yaitu *"feature"*.

Selanjutnya, proses dilakukan *split data* pada *dataframe timeseriesDF*. Dan juga ditentukan *parameter grid* untuk *tuning* yang digunakan dan juga melatih model agar efisiensi dan proses pengolahan data berlangsung secara cepat. Selanjutnya dilakukan evaluasi model dengan menggunakan *evaluator* dan model dilatih menggunakan *Crossvalidator*. Dan diperoleh nilai parameter terbaik pada model yang dibangun yang berupa nilai *maxdepth best model*, nilai *featuresubsetstrategy of best* model serta nilai *Root Mean Square Error (RMSE)*.

Setelah model dibangun selanjutnya diprediksi, hasil prediksi model *Gradient Boosted Tree* divisualisasikan menggunakan l*ibrary matplotlib* didalam *environment pyspark* . Dari hasil visualisasi model *Gradient Boosted Tree* kita bisa mengetahui nilai prediksi total korban kecelakaan pada tahun akan datang yang ditunjukkan pada garis line prediksi pada visualisasi model *Gradient Boosted Tree*. Sehingga setelah dilakukan pengolahan analisis *big data* dengan mengimplementasikan *pyspark* dengan model *K-Means* dan *Gradient Boosted Tree* diperoleh hasil yang kluster kecelakaan serta prediksi total korban kecelakaan pada tahun akan datang .

d. *Flow Diagram*



Gambar 1. Alur Penelitian

Hasil dan Pembahasan

Setelah dilakukan pemrosesan yang telah dilakukan berikut adalah hasil yang telah diperoleh :

Tabel *Hotspot Analysis* yang mengidentifikasi sektor-sektor rawan kecelakaan dengan frekuensi kecelakaan kendaraan paling sering, sehingga memungkinkan dilakukannya intervensi yang ditargetkan pada lokasi 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 kebijakan, seperti kebijakan terkait keselamatan jalan raya, pembuatan peraturan lalu lintas yang lebih ketat,

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

peningkatan infrastruktur, atau kampanye kesadaran masyarakat di kota-kota yang diidentifikasi. terlihat pada Gambar 2.

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

Gambar 2. Identifikasi kecelakaan berdasarkan wilayah

Tabel Hubungan antara Penyebab utama dengan Kendaraan Utama menampilkan informasi mengenai faktor penyebab Utama dan kendaraan utama yang menyebabkan yang diurutkan dari yang paling sering dan ditampilkan 5 teratas. Diinterpretasikan bahwa jenis kendaraan *passenger vehicle* dan sedan dengan *factor unspecified* paling sering kemudian disusul oleh sedan dan *station wagon* dengan faktor lalai dalam berkendara sehingga dapat dijadikan sebagai acuan dalam pembuatan peraturan. terlihat pada Gambar 3.

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

Gambar 3. Hubungan antara penyebab utama dengan kendaraan utama

Selanjutnya, terdapat informasi mengenai jenis kecelakaan yang telah terjadi dimana faktor *Unspecified* atau faktor yang tidak diketahui merupakan jenis kecelakaan yang sering terjadi pada negara *United States* ini dimana total terjadinya hingga 710,122. Lalu pada tabel di informasikan jenis kecelakaan kedua yang sering terjadi adalah faktor dari pengemudi yang terdistraksi.

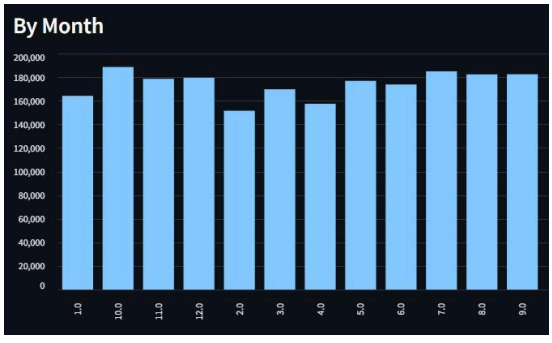
Causes of Accidents

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

Gambar 4. Identifikasi kecelakaan berdasarkan wilayah

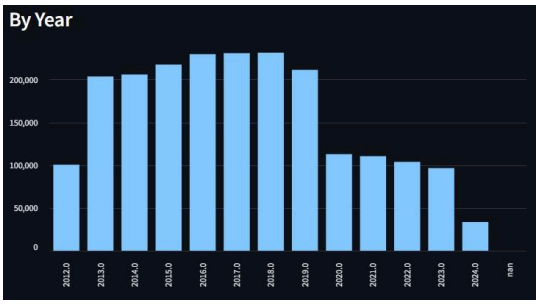
Untuk melihat total kecelakaan motor yang terjadi di *United States* dalam rentang Bulan, dapat dilihat pada bulan ke tujuh dan sepuluh yaitu tepatnya bulan juli dan oktober tingkat kejadian kecelakaan kendaraan bermotor di taraf yang ditunjukan hingga mencapai 185,000 kecelakaan terjadi. Lalu untuk taraf pada bulan Mei, November dan Desember taraf kecelakaan yang telah terjadi menunjukkan ke angka 180,000. Dan untuk taraf terendah ada pada bulan februari dan april yang hanya mencapai ke angka 145,000 terjadinya kecelakaan mengendarai sepeda motor. terlihat pada Gambar 5.



Gambar 5. Tingkat kecelakaan berdasarkan bulan

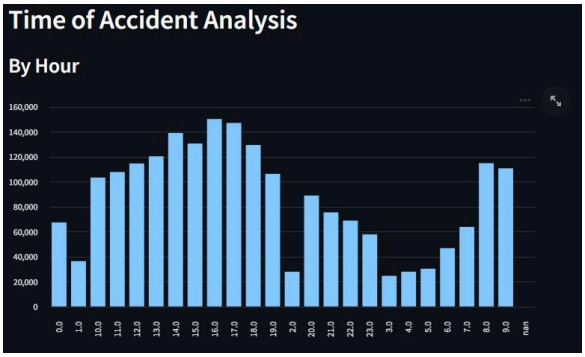
Grafik ini memungkinkan untuk mengamati tren kecelakaan kendaraan secara keseluruhan selama bertahun-tahun. Dapat digunakan untuk menilai efektivitas langkah-langkah keselamatan kemudian evaluasi tentang peraturan yang lebih ketat atau infrastruktur jalan yang lebih baik, dan

Identifikasi potensi penyebab fluktuasi dari waktu ke waktu.. terlihat pada Gambar 6.



Gambar 6. tingkat kecelakaan berdasarkan Tahun

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 Gambar 7.



Gambar 7. Tingkat kecelakaan berdasarkan waktu

Hasil *clustering* yang dilakukan pada data kecelakaan berdasarkan jumlah korban terluka (Total Injured) dan jumlah korban tewas (Total Killed) dengan Tiga klaster yang berbeda diidentifikasi dengan warna yang berbeda:

Kluster 0 (Warna Terang/Pink Muda):

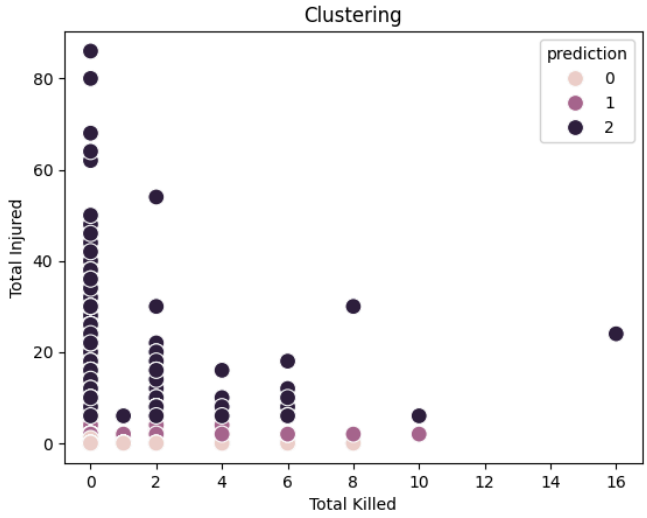
Kluster ini merepresentasikan kecelakaan dengan dampak ringan. Dengan karakteristik didominasi data yang memiliki jumlah korban tewas yang rendah (0-2) dan korban terluka yang rendah (0-4). Ini menunjukkan bahwa kecelakaan dalam kelompok ini cenderung tidak terlalu parah.

Kluster 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.

Kluster 2 (Warna Gelap/Ungu Gelap):

Klaster 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. terlihat pada Gambar 8.

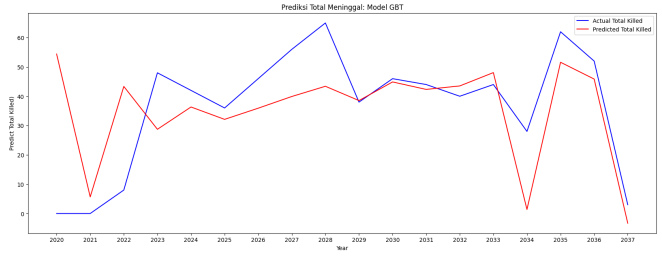


Gambar 8. *Clustering* pada data kecelakaan

Dengan menganalisis dan menginterpretasikan visualisasi ini, dapat diperoleh wawasan mengenai pola, tren, dan penyebab utama kecelakaan kendaraan. Informasi ini kemudian dapat digunakan untuk menginformasikan strategi keselamatan yang efektif, keputusan kebijakan, dan alokasi sumber daya untuk mengurangi frekuensi dan tingkat keparahan kecelakaan serta meningkatkan keselamatan jalan raya secara keseluruhan. Visualisasi ini

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

juga membantu dalam memahami distribusi dan tingkat keparahan kecelakaan yang terjadi. Hal ini bisa berguna untuk analisis lebih lanjut mengenai mengambil langkah pencegahan yang lebih efektif berdasarkan tingkat keparahan kecelakaan. terlihat pada Gambar 9.



Gambar 9. Interpretasi visualisasi

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.

Konflik Kepentingan

Tidak ada konflik kepentingan dalam penelitian ini.

Ucapan Terima Kasih

Penulis menyampaikan terimakasih kepada Dosen Mata Kuliah Analisis Big Data Program Studi Sains Data ITERA atas pemberian Tugas Besar Analisis Big Data yang bertujuan melatih dalam memahami big data.

Referensi

[1] World Health Organization (WHO).(2018). “Laporan Status Global Keselamatan Jalan 2018”

[2] World Health Organization (WHO).(2009). “Time For Action”. Department of Injuries and Violence Prevention Noncommunicable Disease and Mental Health Cluster Disease World Health Organization. Geneva

[3] Sutarto.(2003). “Pengaruh Pemakaian Helm dan Kecepatan Kendaraan Terhadap Tingkat Beratnya Trauma Akibat Kecelakaan Lalu Lintas Pada Pengemudi Sepeda Motor”. Tesis, Universitas Diponegoro, Semarang.

[4] lamtrakul, Pawine.(2003). “Analysis of Motorcycle Accident in Developing Countries: A Case Study of Khon Kaen”. Thailand, Journal of the Eastern Asia Society for Transportation Studies, Vol.5, www.eats.info, diakses tgl. 22 Mei 2024 Hlm. 147-162.

[5] Wedaga, D.M.Priyantha.(210). “Analysing Motorcyle Injuries on Arterial Roads in Bali Using Multinomial Logit Model”. Journal of Eastern Asia Society for Transportation Studies, Vol.8, www.easts.info, diakses tgl. 22 Mei 2024. Hlm.1892-1904.

[6] US Derpatement of Transportation (DOT).(2007). “Motorcylce Safety: A Compilation of Research”. Washington, DC: National Highway Traffic Safety Administration.

[7] Ahmed, S., Islam, M.S., & Habib, M.A. (2008). “An Analysis of Motor Vehicle Accident in the United States Using K-Means Clustering”. Springer Science Business Media

[8] Wang, H., Sun, Y., & MA, X. (2019). “K-Means Clustering for aUnited States”. Transportation Safety and Environment

[9] Smith, J., Jones, M., & Williams, D. (2022). “Using K-Means Clustering to Anlyze Motor Vehicle Accident Data in the United States: A Case Study”. Springer Science Business Media

Lampiran

Pseudocode

1. K-Means

Input:
D= {t1, t2, …. Tn } // Set of elements
K // Number of desired clusters
Output:
K // Set of clusters
K-Means algorithm:
Assign initial values for *m1, m2,…. mk*
repeat
assign each item ti to the clusters which has the closest mean;
calculate new mean for each cluster;
until convergence criteria is met;

- GBT

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.

2. Code

https://colab.research.google.com/drive/12SqA252OeV8P9hxiEzSLm8R6TFj0VneH?usp=sharing (K-Means)

https://colab.research.google.com/drive/1Cv3Y3Gri7rkXz8TjvoQCqP4GuZwoneOg?usp=sharing (GBT)

3. Dataset

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

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia