**DSLS Mini Project Report**

**Data Science**

Study Case: Traffic Jams Prediction on Street Level

Gatum Erlangga

# **Business Understanding**

Tingginya tingkat kepemilikan kendaraan bermotor di Indonesia berpengaruh juga pada peningkatan jumlah pengguna jalan. Hal ini dapat menyebabkan kemacetan dan penumpukan arus pengguna kendaraan bermotor. Kemacetan dapan berdampak pada bisnis seperti transportasi umum, pengiriman dan distribusi barang. Dengan masalah yang ada ini, perlukan sebuah cara untuk alternatif jalan agar dapat terhindar dari kemacetan. Memprediksi tingkat kemacetan jalan sebelum kemacetan terjadi dapat digunakan sebagai solusi. Tingkat kemacetan jalan akan berdampak pada weekend, liburan, jam sibuk dan sebagainya. Dengan membuat model machine learning yang baik, diharapkan pengguna dapat menghemat waktu dan biaya untuk bisa terhindar dari kemacetan.

# **Data Understanding**

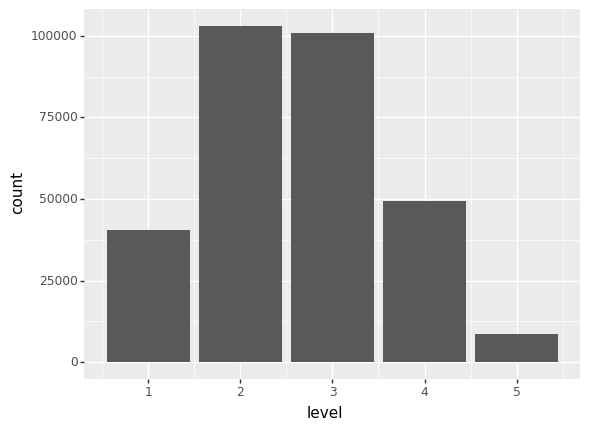
Dataset hanya akan berasal dari Waze User-Generated Data yang telah diproses. Dataset yang akan digunakan adalah [aggregate\_median\_jams](https://drive.google.com/file/d/1N4iWG3SXdBoCDFIIUPt5urebCZC__L32/view?usp=share_link) atau hasil agregasi report pengguna dari level kemacetan di daerah tertentu di kota bandung.

**Tabel 1. Kamus data**

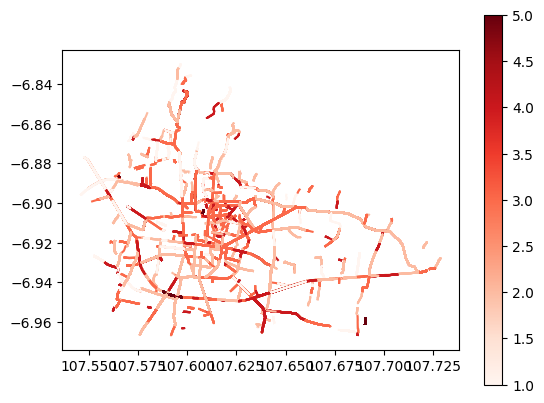
| **Field Name** | **Type** | **Description** | **Value Example** |
| --- | --- | --- | --- |
| time | timestamp | Time (every hour) | 2020-01-13 9:00:00 |
| kemendagri\_kabupaten\_kode | varchar | Kemendagri code of cities | 32,71 |
| kemendagri\_kabupaten\_nama | varchar | Kemendagri name of cities | KOTA BOGOR |
| street | text | Street name | N11 Raden KH Abdullah bin Nuh |
| level | int | Traffic congestion level (median) | 5 |
| median\_length | float | Jam length in meters (median) | 17,6 |
| median\_delay | float | Delay of jam (in seconds) compared to free flow speed (in case of block, -1) (median) | 11,2 |
| median\_speed\_kmh | float | Current median speed on jammed segments in km/h | 78,8 |
| total\_records | int | Total data recorded in a given time | 19 |
| id | serial | Row ID | 569.342 |
| date | date | Date | 2020-01-13 |
| geometry | geometry | Geometry data type (spatial data) | MULTILINESTRING((.., ..)) |

Level 5 memiliki jumlah data yang lebih sedikit dibandingkan dengan level lainnya

Level Distribution



Contoh plot kemacetan (level) pada setiap jalan di kota bandung berdasarkan fitur geometry

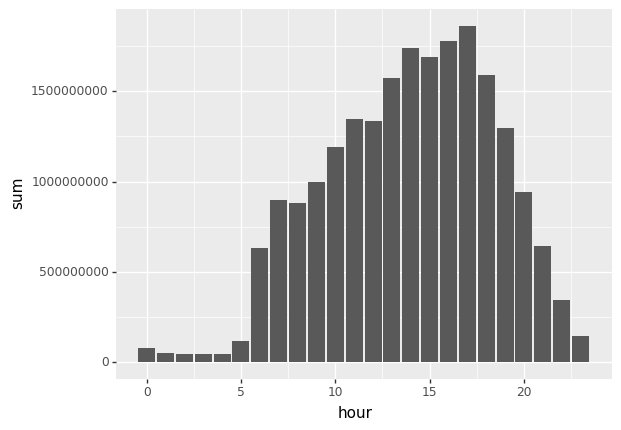


# 

# **Data Cleansing and Preprocessing**

1. **Weekend dan hari libur**
2. **Time column (datetime type) to hour only**

Hour Distribution based on report counts

****

1. **Peak hour and busy hour**

Peak hour: hour >= 13 and <= 18

Busy hour: hour >= 6 and <= 23

1. **Holiday total records metrics.**

The sum of total records, mean, q25, q75, q50.

1. **Weekend total records metrics.**

The sum of total\_records, mean, q25, q75, q50.

1. **Street level features: Street + weekend**

Feature aggregation of total\_records and level with metrics q25, q50/median, q75, average, sum.

1. **Street level features: Street + holiday**

Feature aggregation of total\_records and level with metrics q25, q50/median, q75, average, sum.

**Final Data**

Data columns (total 46 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 time 301995 non-null object

1 kemendagri\_kabupaten\_kode 301995 non-null float64

2 kemendagri\_kabupaten\_nama 301995 non-null object

3 street 299552 non-null object

4 level 301995 non-null int64

5 median\_length 301995 non-null float64

6 median\_delay 301995 non-null float64

7 median\_speed\_kmh 301995 non-null float64

8 total\_records 301995 non-null int64

9 id 301995 non-null int64

10 date 301995 non-null object

11 median\_level 301995 non-null float64

12 geometry 301995 non-null object

13 thm 301995 non-null object

14 is\_weekend 301995 non-null int64

15 is\_holiday 301995 non-null int64

16 hour 301995 non-null int64

17 peak\_hour 301995 non-null int64

18 busy\_hour 301995 non-null int64

19 total\_records\_holiday\_sum 301995 non-null int64

...

44 level\_q50\_y 301995 non-null float64

45 level\_q75\_y 301995 non-null float64

dtypes: float64(28), int64(9), object(9)

memory usage: 108.3+ MB

# 

# **Modeling**

Cols X: ['street', 'median\_length', 'median\_delay', 'median\_speed\_kmh', 'total\_records', 'median\_level', 'is\_weekend', 'is\_holiday', 'hour', 'peak\_hour', 'total\_records\_holiday\_sum', 'total\_records\_holiday\_q25', 'total\_records\_holiday\_q75', 'total\_records\_holiday\_mean', 'total\_records\_holiday\_q50', 'total\_records\_busy\_hour\_mean', 'total\_records\_busy\_hour\_sum', 'total\_records\_busy\_hour\_q25', 'total\_records\_busy\_hour\_q50', 'total\_records\_busy\_hour\_q75', 'total\_records\_avg\_x', 'total\_records\_q25\_x', 'total\_records\_q50\_x', 'total\_records\_q75\_x', 'level\_q25\_x', 'level\_q50\_x', 'level\_q75\_x', 'total\_records\_avg\_y', 'total\_records\_q25\_y', 'total\_records\_q50\_y', 'total\_records\_q75\_y', 'level\_q25\_y', 'level\_q50\_y', 'level\_q75\_y']

Col target (y) = [‘level’]

Library: Sklearn

Model: Random Forest

Parameter: max\_depth=10, n\_estimators=35, random\_state=42

Steps:

1. Undersampling to reduce bias
2. Street label encoder
3. Train test split (test=0.33)
4. Model building with parameter
5. Model training
6. Evaluation

# **Evaluation**

metrics.accuracy\_score(y\_pred,y\_test)

0.9986052438816364

metrics.confusion\_matrix(y\_test, y\_pred)

array([[13060, 98, 4, 0, 0],

[ 0, 33946, 0, 0, 0],

[ 0, 0, 33348, 0, 0],

[ 0, 0, 37, 16336, 0],

[ 0, 0, 0, 0, 2830]], dtype=int64)

# 

# **Referensi**

[Undersampling Algorithms for Imbalanced Classification](https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/)

[Big Data Analysis and Prediction of Traffic in Los Angeles](https://www.koreascience.or.kr/article/JAKO202011161036290.page)