PREDICTIVE ANALYTICS MENGGUNAKAN *MACHINE LEARNING* UNTUK MEMPREDIKSI WAKTU KETERLAMBATAN BERDASARKAN PENYEBAB KETERLAMBATAN PADA PT. KERETA API INDONESIA

Christopher Sanjaya[1], Suhono Harso Supangkat[2]
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
Bandung, Indonesia
sanjayachristopher46@gmail.com[1], suhono@stei.itb.ac.id[2]

Abstract - PT. Kereta Api Indonesia (KAI) is a company that regulates trains in Indonesia. Railways in Indonesia still often experience delays, especially in the Cibatu Purwakarta lane which will be the object of research in this study. This research is intended as an initial stage of applying machine learning to overcome the problem of tardiness by providing the best model for predicting tardiness and what things are causing the tardiness pattern. Machine learning models considered are decision tree regression, support vector machine regression, random forest regression, ensemble learning, and gradient boosting regression. From the best machine learning techniques, a model will be made to predict the delay based on the cause of the delay.

Keywords – KAI, machine learning, predict the delay, cause of delay, Cibatu Purwakarta lane

1. Pendahuluan

Kereta api adalah bentuk transportasi rel yang terdiri dari serangkaian kendaraan yang ditarik sepanjang jalur kereta api untuk mengangkut kargo atau penumpang. Kereta api bisa terdiri dari kombinasi satu atau lebih lokomotif dan gerbong kereta. Terdapat dua jenis kereta api dari segi pengunaannya, yaitu kereta api penumpang dan kereta api barang. Kereta api penumpang adalah satu rangkaian kereta lokomotif yang digunakan untuk mengangkut manusia. Kereta api barang adalah kereta api yang digunakan untuk mengangkut barang (kargo).

Di Indonesia kereta api diatur oleh PT Kereta API Indonesia atau yang lebih sering disingkat menjadi KAI atau PT KAI. Penggunaan kereta api semakin meningkat, Pada tahun 2018 terdapat 428 juta penumpang, sedangkan tahun sebelumnya sebesar 394 juta penumpang. Dapat dilihat peningkatan penggunaan kereta api dari tahun sebelumnya meningkat sebanyak sepuluh persen.

Keterlambatan kereta api merupakan suatu yang sering terjadi pada PT Kereta Api Indonesia. Keterlambatan kereta api menjadi salah satu masalah utama yang perlu diselesaikan agar KAI sesuai dengan salah satu pilar utama KAI, yaitu ketepatan waktu. Keterlambatan dapat menyebabkan terlambatnya penumpang ke suatu tempat tujuan dan dapat mengganggu aktivitas yang sudah direncanakan penumpang.Suatu kereta diangga terlambat apabila memiliki keterlambatan sebesar 5 menit.Keterlambatan kereta api penumpang di Indonesia rata – rata adalah 10 - 15 menit pada tahun 2017. Terdapat juga banyak kejadian dimana terdapat keterlambatan kereta sampai lebih dari setengah jam, KA Logawa tujuan Stasiun Lempuyangan, Kereta Bengawan tujuan Stasiun Pasar Senen.

Terdapat berbagai macam penyebab keterlambatan, yaitu pengurangan kecepatan akibat ada rel kereta yang sedang diperbaiki, matinya listrik untuk menjalankan kereta api, anjloknya kereta api akibat kondisi rel kereta api tidak baik, dan pembangunan jalur rel kereta api baru.

PT. KAI melakukan prediksi keterlambatan dengan menggunakan GPS dan perhitungan jarak dibagi kecepatan kereta api normal sehingga kurang akurat apabila terjadi keterlambatan di banyak lokasi dikarenakan hanya menghitung keterlambatan awal saja. Oleh sebab itu, untuk meningkatkan akurasi prediksi keterlambatan digunakan konsep machine learning. Machine learning digunakan untuk menyelesaikan masalah yang memiliki solusi yang harus terus dilakukan *update* secara manual atau memiliki daftar aturan yang banyak dan mendapatkan pengetahuan mengenai suatu permasalahan yang kompleks dan jumlah data yang banyak [6]. Penelitian ini merupakan tahap awal dari implementasi *machine learning* pada KAI. Diharapkan penelitian ini dapat menambahkan wawasan dan meningkatkan pelayanan KAI mengenai masalah keterlambatan.

Untuk mencari solusi mengenai keterlambatan kereta, penelitian ini akan membahas penggunaan *machine learning* untuk membuat model prediksi keterlambatan kereta berdasarkan jenis keterlambatannya. *Machine learning* yang akan dibahas adalah dengan menggunakan *decision tree regression [1], support vector machine regression [6], random forest regression [2], ensemble learning [6],* dan *gradient boosting regressor [3]*. Akan dicari model terbaik untuk memprediksi keterlambatan pada KAI.

Berdasarkan latar belakang yang sudah dijabarkan sebelumnya, penelitian ini membahas mengenai algoritma *machine learning* yang digunakan untuk memprediksi keterlambatan berdasarkan jenis keterlambatan pada perusahaan Kereta Api Indonesia. Berikut *detail* rumusan masalah yang dibahas dalam penelitian.

1. Bagaimana algoritma *machine learning* dapat membuat model yang memiliki akurasi prediksi yang tinggi terhadap jenis keterlambatan di PT. KAI?
2. Bagaimana penerapan algoritma rancangan *machine learning* dalam menambahkan wawasan terhadap PT. KAI?

Berikut merupakan ruang lingkup yang dibahas dalam penelitian.

1. Hasil akhir pada penelitian berupa model dari penerapan algortima *machine learning* dan kesimpulan dri penerapan tersebut
2. Data yang digunakan merupakan jadwal keberangkatan dan kedatangan kereta api di stasiun kereta api Bandung jalur Cibatu Purwakarta dan jalur Purwakarta Cibatu dari Januari 2019 – Maret 2019
3. Prediksi keterlambatan dilakukan dengan *input* tanggal keterlambatan, jenis keterlambatan, akibat keterlambatan, daerah operasi, waktu keterlambatan, lokasi, hari ke, bulan ke, minggu ke, dan nama hari
4. Pemodelan algoritma *machine learning* menggunakan bahasa pemrograman *Python*

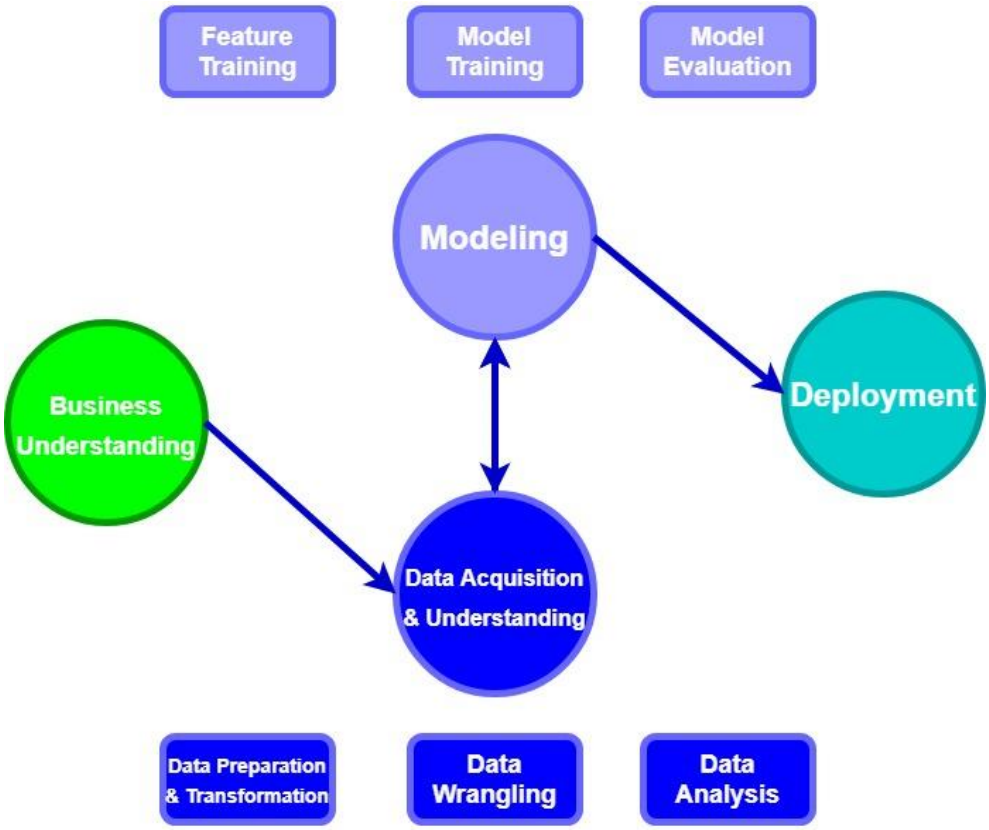
Ada beberapa penelitian mengenai keterlambatan kereta yang serupa dengan penelitian yang dilakukan. *Predictions of train delays using machine learning* oleh Robert Nillson dan Kim Henning, pada tahun 2018 [1]. Penelitian tersebut dilakukan terhadap keterlambatan kereta di Stockholm. Prediksi keterlambatan dilakukan dengan menggunakan *decision tree, adaBoost,* dan *neural network.*

Penelitian selanjutnya adalah oleh Ramashish Gaurav dan Biplav Sricastava yang memiliki judul *Estimating Train Delays in a Large Rail Network Using Zero Shot Markov Model* pada tahun 2018 [2]. Penelitian tersebut dilakukan terhadap keterlambatan kereta di India. Prediksi keterlambatan dilakukan dengan menggunakan data historis dan data ramalan cuaca.Penelitian tersebut mengunakan *Markov framework* dan menggunakan *random forest regression* dan *ridge regression* untuk memprediksi keterlambatan.

Penelitian yang lain adalah *Train delay analysis and prediction based on big data fusion* yang ditulis oleh Pu Wang dan Qing-peng Zhamg pada tahun 2019 [3]. Tujuan dari penelitian ini adalah menggabungkan data historis dan data ramalan cuaca untuk memeprediksi keterlambatan pada setiap stasiun di India.Penelitian tersebut juga mencari nilai keterlambatan untuk menentukan propagasi keterlambatan antar stasiun. Penelitian tersebut menggunakan *density based clustering algorithm* dan *gradient boosted regression trees* untuk melakukan prediksi.

2. Tahapan Pembuatan Program

Berikut ini merupakan deskripsi tahapanmodeldari pengembangan model yang akan dilakukan. Alur tahapan menggunakan *The Team Data Science Process lifecycle [4]*.



Gambar 1 ***The Team Data Science Process lifecycle***

A. Business Understanding

Pada tahapan ini dilakukan diskusi dengan Manager Operasional KAI untuk mengetahui permasalahan yang dihadapi pada saat ini. Setelah melakukan riset dan diskusi, dirumuskan

bahwa KAI dapat diuntungkan dengan dibuatnya sebuah model prediksi keterlambatan menurut penyebab keterlambatan. Dengan adanya solusi tersebut, diharapkan KAI dapat mendapatkan wawasan mengenai setiap jenis keterlambatan dan pola keterlambatan.

B. Data Acquisition & Understanding

Pada tahap ini dilakukan pengumpulan data yang diperlukan untuk melakukan prediksi. Setelah data dikumpulkan, akan dilakukan *filter*ing dan *cleaning* terhadap data. Pada proses *filter*ing dilakukan pembagian data ke dalam bentuk yang lebih kecil untuk memahami gambaran besar dari data tersebut. Pada proses cleaning dilakukan pembersihan terhadap data yang memiliki missing value atau data yang tidak lengkap. Pada tahap ini juga dilakukan penghapusan terhadap tabel yang dianggap tidak diperlukan.Untuk memahami dan mendapatkan bentuk data yang diperlukan untuk melakukan prediksi, dilakukan beberapa tahap, yaitu *data preparation* & *transformation*, *data wrangling*, dan *data analysis*.

Tabel 1 Data Mentah

No	Nama Atribut	Jenis Data
1	Tanggal	Date
1	Penyebab	String
2	Akibat	String
3	Daop	Integer
4	Lokasi	String
5	Lokasi	String
6	KM	String
7	Andil	Integer
8	Keterangan	String

Pada Tabel 1 merupakan bentuk data mentah yang didapat dari KAI. Tanggal merupakan tanggal suatu keterlambatan terjadi, Penyebab merupakan penyebab suatu keterlambatan terjadi, Akibat merupakan akibat dari suatu keterlambatan, Daop merupakan daerah operasi kereta tersebut, Lokasi merupakan lokasi awal keterlambatan terjadi, Lokasi merupakan lokasi akhir suatu keterlambatan, KM merupakan jarak perjalanan dari stasiun awal dan pada KM berapa suatu keterlambatan terjadi, Andil merupakan waktu keterlambatan yang terjadi, Keterangan merupakan keterangan lebih rinci mengenai keterlambatan yang terjadi.

Tabel 2 Jumlah Data yang Diteliti

No	Jenis Data	Jumlah Data
1	Jumlah Kereta yang diteliti	2
2	Jumlah Lokasi yang diteliti	29
3	Jumlah data keterlambatan yang didapat	707
4	Jumlah data (dalam bulan)	2

Pada Tabel 2 dijelaskan bahwa jumlah kereta yang diteliti adalah 2 kereta, jumlah lokasi yang diteliti adalah 29 lokasi, jumlah data keterlambatan berdasarkan jenis keterlambatan yang didapat adalah 707 data, data yang diteliti adalah 2 bulan. No unik kereta yang diteliti adalah nomor 395 dan nomor 396. Lokasi yang diteliti adalah Cibatu, Leuwigoong, Karangsari, Leles, Lebakjero,

Nagreg, Cicalengka, Haurpugur, Rancaekek, Cimekar, Gedebage, Kiaracondong, Cikudapateuh, Bandung, Andir, Cimindi, Cimahi, Gadongbangkong, Padalarang, Cilame, Sasaksaat, Maswati, Rendeh, Cikadongdong, Cisomang, Plered, Sukatani, Ciganea, dan Purwakarta, gambar rute dapat dilihat pada Gambar 2. Data yang diteliti adalah bulan Februari 2019 sampai bulan Maret 2019.

Gambar 2 Rute Kereta Api Lokal Cibatu

C. Data preparation & Transformation

Pada tahap ini dilakukan pengubahan terhadap data ke dalam bentuk csv agar dapat diproses oleh *Jupyter Notebook.* Dilakukan pula pengubahan terhadap data – data string agar dapat di proses dengan *One Hot Encoder* yang terdapat pada *library* sklearn. *One Hot Encoder* akan membuat data string atau data kategori menjadi *sparse matrix. Sparse matrix* adalah sebuah matrix yang memiliki banyak nilai nol di dalamnya. Data – data yang memiliki *missing value* juga ditangani dengan menghapus baris tersebut.

Tabel 3 ***Output Data Preparation & Transformation***

No	Nama Atribut	Jenis Data
1	Tanggal	Date
2	Penyebab	String
3	Akibat	String
4	Lokasi_1	String
5	Lokasi_2	String
6	Andil	Integer
7	No Kereta	Integer
8	Bulan Ke	Integer
9	Hari_Ke	Integer
10	Nama Hari	Integer
11	Minggu Ke	Integer
12	Andil2	Integer
13	Andil3	Integer

D. Data wrangling

Pada tahap ini dilakukan penghapusan terhadap tabel data yang dianggap tidak diperlukan.Pada tahapan ini pula dilakukan pemetaan setiap variabel dalam analisis yang memiliki keterhubungan. Data yang dianggap relevan satu dengan lainnya dibuat ke dalam satu tabel. Kolom yang dibuang adalah kolom Tanggal karena sudah diubah menjadi bentuk integer, kolom Hari Ke karena tidak memiliki pola, kolom Andil2 dan Andil3 karena kolom tersebut digunakan untuk penghitungan rata – rata saja. Output untuk data wrangling dapat dilihat pada Tabel 4.

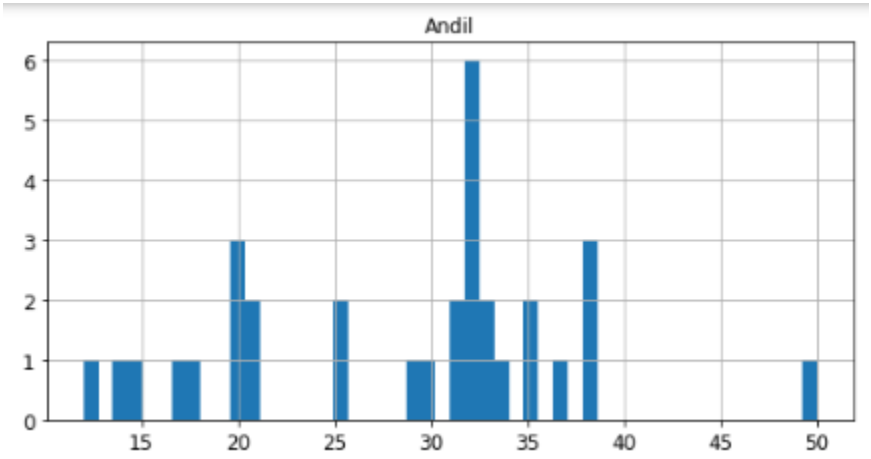
Table 4 ***Output Data Wrangling***

No	Nama Atribut	Jenis Data
1	Penyebab	String
2	Akibat	String
3	Lokasi_1	String
4	Lokasi_2	String
5	Andil	Integer
6	No Kereta	Integer
7	Bulan Ke	Integer
8	Nama Hari	Integer
9	Minggu Ke	Integer

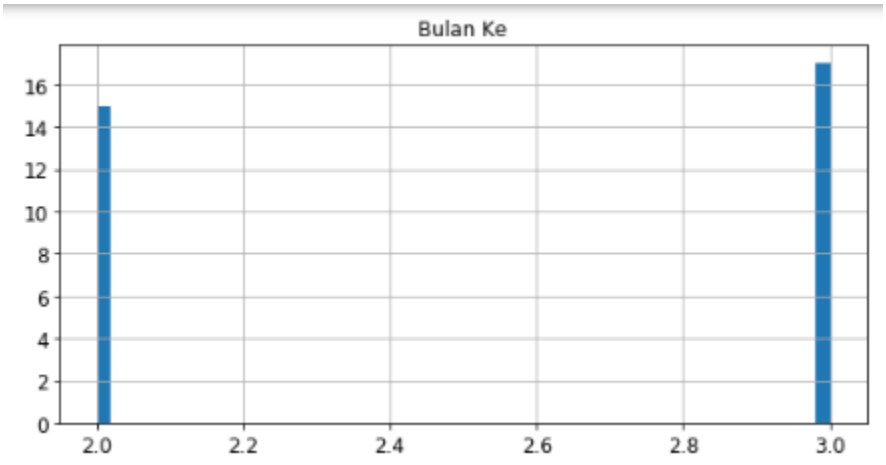
E. Data analysis

Pada tahapan ini dilakukan pembuatan histogram untuk melihat keterhubungan data yang satu dengan data yang lainnya. Data – data yang sudah dipetakan dicari keterhubungannya satu dengan lainnya. Dicari tahu juga gambaran dari setiap histogram yang ada dan makna dari diagram tersebut.

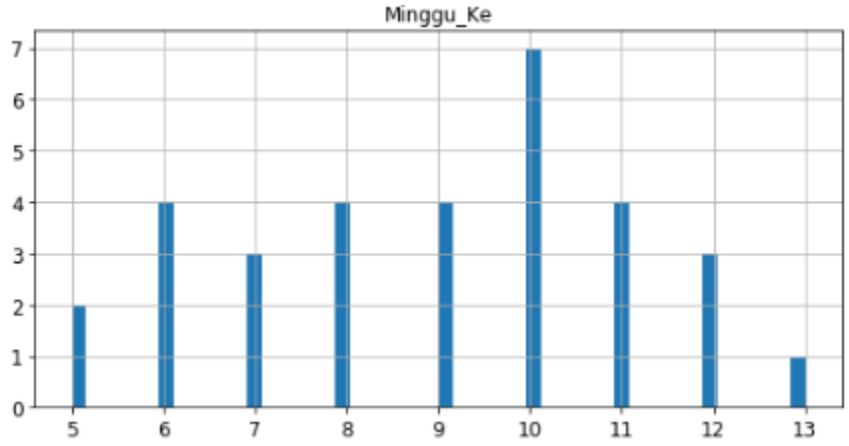
Pada Gambar 3 dapat dilihat bahwa Andil untuk nomor 395 berkisar antara 10 menit – 50 menit. Pada Gambar 4 dapat dilihat bahwa pada bulan Februari, terhadap 15 kasus keterlambatan sedangkan pada bulan Maret terdapat 17 kasus keterlambatan. Pada Gambar 5 dapat dilihat bahwa keterlambatan pada minggu 10 terjadi setiap hari sedangkan untuk minggu lainnya berkisar antara 1 hari sampai 4 hari. Pada Gambar 6 dapat dilihat bahwa keterlambatan pada umumnya terjadi pada hari Jumat dan Sabtu dibandingkan dengan hari lainnya (0=Senin).



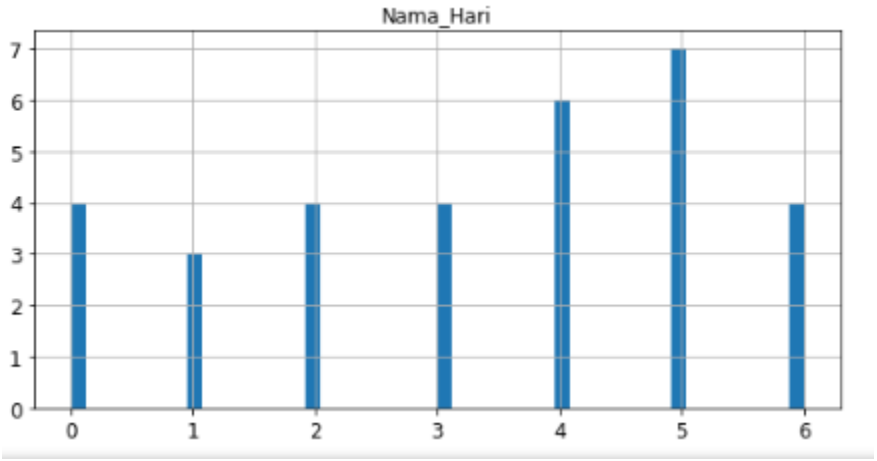
Gambar 3 Andil Kereta 395



Gambar 4 Jumlah Keterlambatan Setiap Bulan Kereta 395

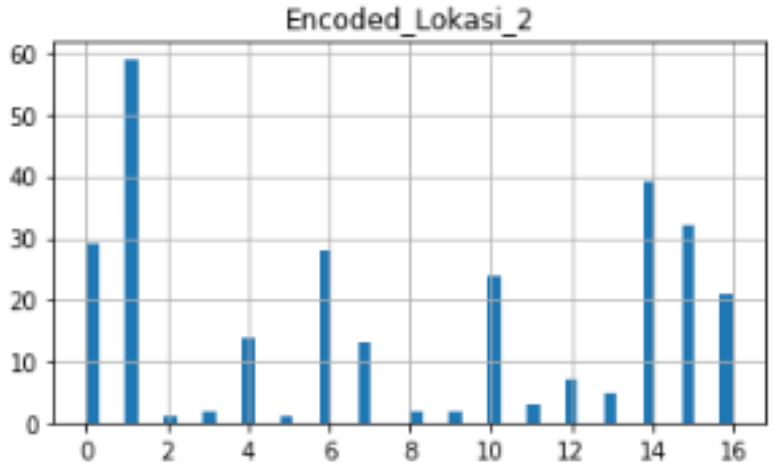


Gambar 5 Jumlah Keterlambatan Setiap Minggu Kereta 395



Gambar 6 Jumlah Keterlambatan Setiap Nama Hari Kereta 395

Pada Gambar 7 dapat dilihat bahwa keterlambatan paling sering terjadi pada nomor 1. Nomor 1 adalah Bandung apabila kita melihat hasil encodingnya pada Gambar 8. Diikuti oleh nomor 14 yaitu Rendeh, nomor 15 Sasaksaat, nomor 0 Andir, nomor 6 Cimekar, nomor 10 Padalarang, nomor 16 Sukatani, nomor 4 Cilame, dan nomor 7 Karangsari untuk keterlambatan yang lebih dari 10 kali. Gambar 8 merupakan hasil encoding dimana 0 adalah Andir, 1 adalah Bandung, dan seterusnya.



Gambar 7 Histogram Keterlambatan Setiap Lokasi Kereta 395

```
1 ordinal_encoder.categories_
```

```
[array(['Andir', 'Bandung', 'Cibatu', 'Cicalengka', 'Cilame', 'Cimahi',
        'Cimekar', 'Karangsari', 'Leles', 'Maswati', 'Padalarang',
        'Plered', 'Purwakarta', 'Rancaekek', 'Rendeh', 'Sasaksaat',
        'Sukatani'], dtype=object)]
```

Gambar 8 Encoding Lokasi Kereta 395

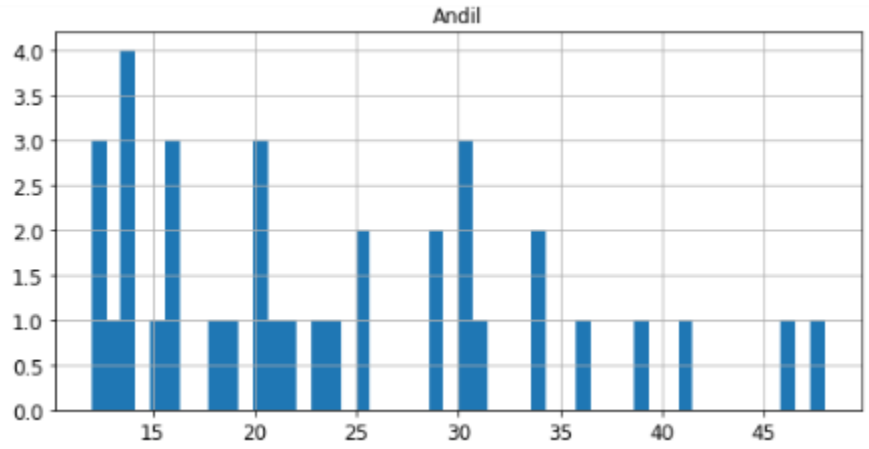
Pada Gambar 9 dapat dilihat bahwa penyebab keterlambatan yang sering terjadi adalah PEMASANGAN TASPAT (PRASARANA), PENAMBAHAN HSD, TUNGGU PERSILANGAN, PEKERJAAN JEMBATAN yang memiliki frekuensi lebih dari 30 kali. TAKTIS MASINIS dan TAKTIS PPKA tidak termasuk ke dalam penyebab keterlambatan karena hal tersebut merupakan pengurangan keterlambatan oleh KAI.

```
1 dt["Penyebab"].value_counts()
```

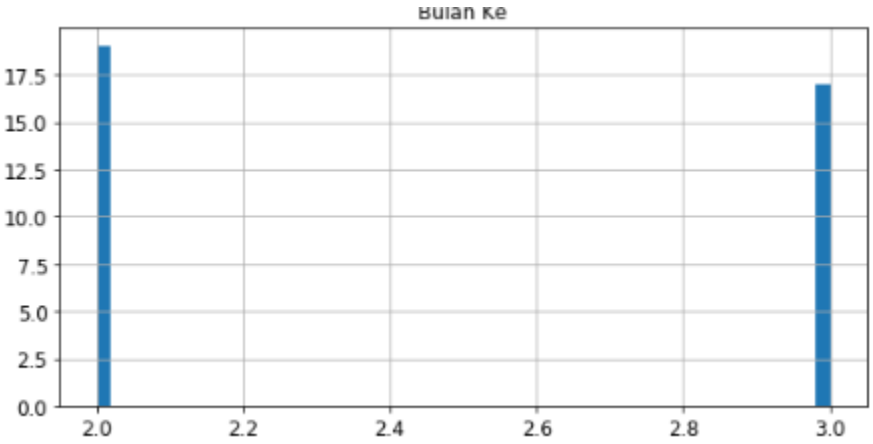
PEMASANGAN TASPAT (PRASARANA)	113
TAKTIS MASINIS	56
PENAMBAHAN HSD KERETA PEMBANGKIT DI STASIUN ANTARA	42
TUNGGU PERSILANGAN	35
PEKERJAAN JEMBATAN	31
TAKTIS PPKA	31
GEOMETRI (JJ)	15
PEMINDAHAN PERSILANGAN	12
TUNGGU PENYUSULAN	7
GARDAN SHAFT	4
TAMBAH/LEPAS LOK TRAKSI GANDA DI STASIUN ANTARA	2
GANGGUAN PENGEREMAN (KERETA)	2
ANJLOGAN (EKSTERNALITAS)	1
TUNGGU RANGKAIAN (ALAM DAN EKSTERNALITAS)	1
GANTI LOK DENGAN LOK CADANGAN DI STASIUN ANTARA	1

Gambar 9 Penyebab Keterlambatan Kereta 395

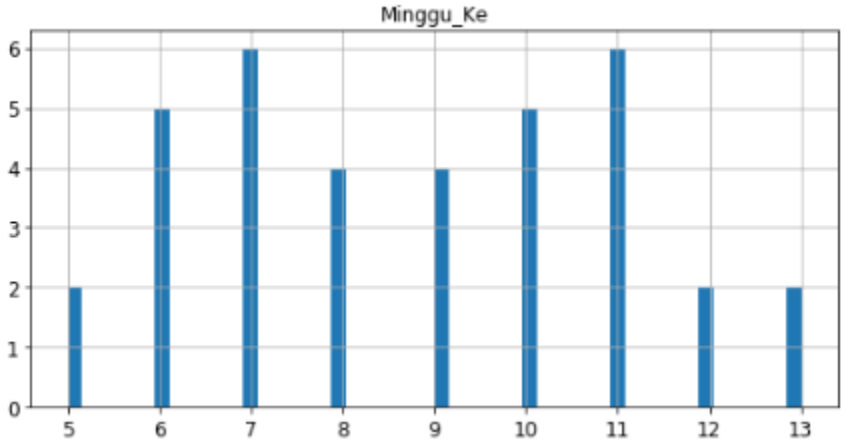
Pada Gambar 10 dapat dilihat bahwa Andil untuk nomor 396 berkisar antara 10 menit – 50 menit.Pada Gambar 11 dapat dilihat bahwa pada bulan Februari, terhadap 18 kasus keterlambatan sedangkan pada bulan Maret terdapat 16 kasus keterlambatan. Pada Gambar 12 dapat dilihat bahwa keterlambatan pada minggu 7 dan 11 terjadi 6 hari sedangkan untuk minggu lainnya berkisar antara 2 hari sampai 5 hari tiap minggunya. Pada Gambar 13 dapat dilihat bahwa keterlambatan pada umumnya terjadi pada hari Senin, tetapi keterlambatan untuk hari lainnya tidak begitu berbeda jauh, hanya pada haru Rabu saja memiliki keterlambatan hanya 2 kali (0=Senin).



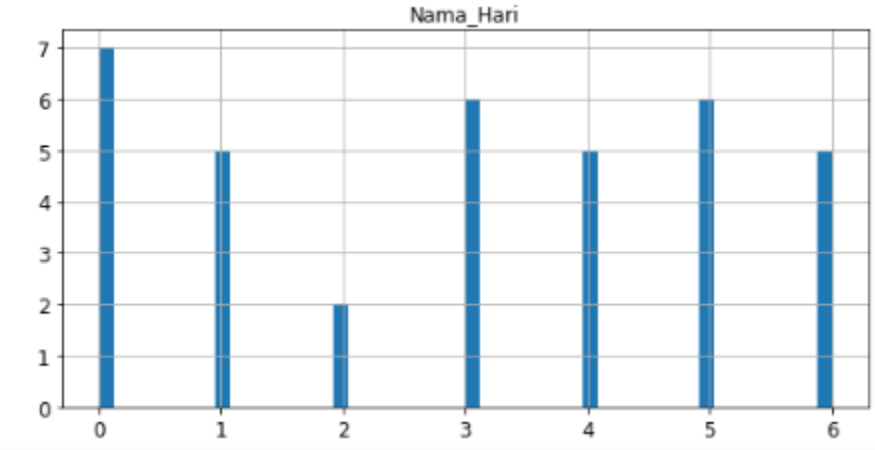
Gambar 10 Andil Kereta 396



Gambar 11 Jumlah Keterlambatan Setiap Bulan Kereta 396

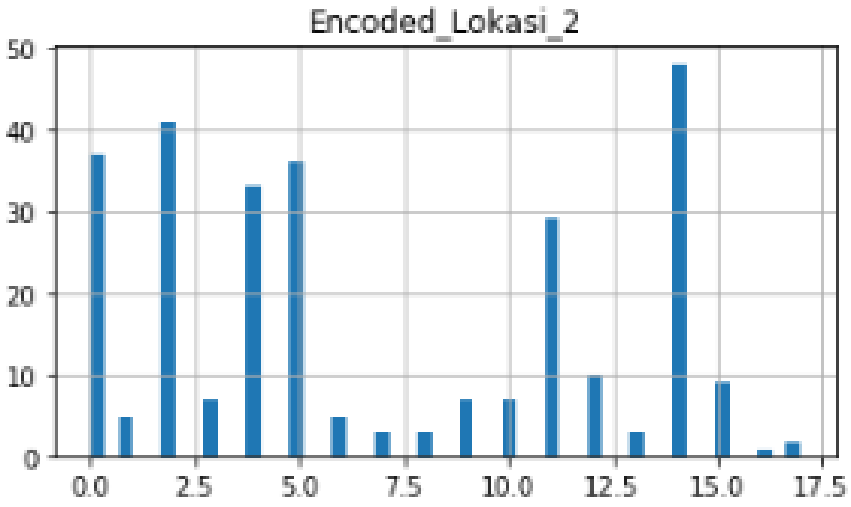


Gambar 12 Jumlah Keterlambatan Setiap Minggu Kereta 396



Gambar 13 Jumlah Keterlambatan Setiap Nama Hari Kereta 396

Pada Gambar 14 dapat dilihat bahwa keterlambatan paling sering terjadi pada nomor 14. Nomor 14 adalah Purwakarta apabila kita melihat hasil encodingnya pada Gambar 15. Diikuti oleh nomor 2 yaitu Cicalengka, nomor 0 Bandung, nomor 5 Cilame, nomor 4 Cikudapateuh, nomor 11 Maswati, dan nomor 10 Nagreg, untuk keterlambatan yang lebih dari 10 kali. Gambar 15 merupakan hasil encoding dimana 0 adalah Bandung, 1 adalah Cibatu, dan seterusnya.



Gambar 14 Histogram Keterlambatan Setiap Lokasi Kereta 396

```
1  ordinal_encoder.categories_
```

```
[array(['Bandung', 'Cibatu', 'Cicalengka', 'Cikadongdong', 'Cikudapateuh',
        'Cilame', 'Cimekar', 'Gedebage', 'Haurpugur', 'Kiaracondong',
        'Leles', 'Maswati', 'Nagreg', 'Purwakarta', 'Rancaekek', 'Rendeh',
        'Sasaksaat', 'Sukatani'], dtype=object)]
```

Gambar 15 Encoded Lokasi Kereta 396

Pada Gambar 16 dapat dilihat bahwa penyebab keterlambatan yang sering terjadi adalah PEMASANGAN TASPAT (PRASARANA), TUNGGU PERSILANGAN, PEKERJAAN JEMBATAN yang memiliki frekuensi lebih dari 30 kali. TAKTIS MASINIS dan TAKTIS PPKA tidak termasuk ke dalam penyebab keterlambatan karena hal tersebut merupakan pengurangan keterlambatan oleh KAI.

```
1  dt["Penyebab"].value_counts()
```

```
PEMASANGAN TASPAT (PRASARANA)                123
TAKTIS MASINIS                                51
TUNGGU PERSILANGAN                            45
TAKTIS PPKA                                   38
PEKERJAAN JEMBATAN                            36
GEOMETRI (JJ)                                 20
TUNGGU PENYUSULAN                             19
ALAM DAN EKSTERNALITAS                         7
PEMINDAHAN PERSILANGAN                         7
ANJLOGAN (EKSTERNALITAS)                       3
GANGGUAN SINYAL (PERSINYALAN ELEKTRIK)         2
TUNGGU RANGKAIAN (ALAM DAN EKSTERNALITAS)      2
ANTRIAN TIKET                                  1
Name: Penyebab, dtype: int64
```

Gambar 16 Penyebab Keterlambatan Kereta 396

F. Modeling

Pada tahapan ini dibuat pembuatan model prediksi yang akan digunakan untuk memprediksi keterlambatan berdasarkan penyebab keterlambatan untuk 2 jalur kereta api, yaitu jalur Cibatu Purwakarta dan jalur Purwakarta Cibatu. Data utama yang digunakan untuk membuat model prediksi adalah data historis keterlambatan kereta api, data penyebab keterlambatan, dan data akibat keterlambatan. Pada tahap ini dilakukan pembuatan model yang memiliki akurasi paling tinggi dari antara pemodelan menggunakan *decision tree regressor, support vector machine, random forest regressor, gradient boosting regressor, ensemble learning*.

G. Feature *Training*

Penentuan fitur utama didapatkan dari penggabungkan penelitian terkait yang sudah dilakukan dan hasil diskusi dari KAI. Dikarenakan fitur yang dimiliki tidak terlalu banyak, maka digunakan hampir seluruh fiturnya, yaitu Penyebab, Akibat, Lokasi_1, Lokasi_2, Bulan_Ke, Minggu_Ke, Nama_Hari [7], No Kereta. Fitur Hari_Ke tidak digunakan karena menyebabkan pengurangan terhadap akurasi yang didapat. Data # yang berupa nomor juga akan dihapus karena hal tersebut tidak ada hubungannya dengan model prediksi.

H. Model *training*

Untuk melakukan pembuatan model prediksi keterlambatan, digunakan Python 3.0 sebagai bahasa pemrograman untuk membuat model prediksi. Digunakan Jupyter Notebook sebagai IDE untuk membuat model prediksi. Dilakukan alokasi data ke dalam data untuk *training* dan data untuk *testing*. Data dibagi ke dalam 80% untuk data *training* dan 20% untuk data *testing*.

Dilakukan pengubahan untuk data dalam bentuk string dengan menggunakan column transformer dan *One hot encoder*. *One hot encoder* akan mengubah data dalam bentuk string ke dalam bentuk biner sehingga akan terdapat matrix dengan jumlah kolom sebanyak jumlah jenis dalam *string* yang akan diubah. Pada data No Kereta terdapat 2 jenis akibat yang menyebabkan keterlambatan kereta api, sehingga *one hot encoder* akan membuat matrix dengan 2 kolom dimana nilai untuk setiap baris akan terdapat nilai 1 dan sisanya 0. Hal tersebut juga dilakukan

terhadap Penyebab, Akibat, Lokasi_1, Lokasi_2, Minggu Ke, Bulan Ke. Setelah dilakukan *one hot encoder*, akan dilakukan penggabungan data dengan data awal agar dapat dilakukan *training*.

Dibuat pelatihan model dengan menggunakan *decision tree regressor, random forest regressor, support vector machine, gradient boosting regressor,* dan *voting regressor*. Penggunaan algoritma pelatihan model tersebut bertujuan untuk mencari algoritma pemodelan terbaik untuk kasus prediksi keterlambatan. GridSearchCV digunakan untuk mencari *hyperparameter* terbaik untuk algoritma *random forest*.

Setelah itu digunakan fitur untuk membuat pipeline yang sudah disediakan oleh sklearn, yaitu make_pipeline. Pipeline digunakan untuk menentukan proses pengerjaan yang dilakukan kernel. Untuk setiap pipeline akan digunakan untuk melatih satu algoritma. Pada pipe digunakan untuk melatih *random forest* regression. Pada pipe2 digunakan untuk melatih support vector machine regression. Pada pipe 3 digunakan untuk melatih *decision tree regressor*. Pada pipe4 digunakan untuk melatih GridSearchCV. Pada pipe5 digunakan untuk melatih *ensemble learning* yang berisikan *random forest regressor*, *support vector machine regressor*, *decision tree regressor*, dan *gradient boosting regressor*.Pada pipe6 digunakan untuk melatih *gradient boosting regressor*. Pipe.fit digunakan untuk melatih model berdasarkan data yang dimasukkan sebagai parameter.

Setelah GridSearchCV dilakukan pelatihan terhadap dataset yang ada, didapatkan bahwa *hyperparameter* untuk max*_features* pada *random forest regressor* adalah 8 dan *hyperparameter* untuk n_estimators pada *random forest regressor* adalah 30.

Setelah seluruh model dilatih, dilakukan penilaian terhadap model yang telah dibuat dengan cross val score. Cross val score melakukan *cross validation* untuk menghitung akurasi yang akan dimiliki oleh model yang dibuat. *Cross validation* disini dilakukan sebanyak 5 kali.

Untuk pelatihan dengan menggunakan *random forest regressor* didapatkan akurasi sebesar 70,7%. Untuk pelatihan dengan menggunakan *support vector machine regressor* didapatkan akurasi sebesar 70,3%. Untuk pelatihan dengan *decision tree regressor* didapatkan akurasi sebesar 60,7%. Untuk pelatihan dengan menggunakan *ensemble learning* didapatkan akurasi sebesar 73,9%. Untuk pelatihan dengan menggunakan gradient boosting regression didapatkan akurasi sebesar 71,8%. Didapatkan kesimpulan bahwa penggunaan *ensemble learning* mendapatkan akurasi paling tinggi sehingga model yang dipakai untuk melakukan prediksi terhadap data testing adalah dengan model *ensemble learning*.

I. Model Evaluation

Untuk cek akurasi, dilakukan tes terhadap model dengan melakukan prediksi terhadap data test yang sudah dipisahkan di awal. Tes akurasi dilakukan dengan menggunakan *mean absolute error* dan bukan *root mean square error* karena *mean absolute error* lebih konsisten *[5]*, yaitu rata – rata perbedaan antara nilai yang diprediksi dengan nilai yang sebenarnya lalu nilai tersebut dibuat absolut agar selalu positif. Model yang dipakai untuk menghitung*mean absolute error* adalah model pipe5, yaitu *ensemble learning*. Nilai yang didapat dari *mean absolute error* adalah 2.21 menit. Berarti rata – rata kesalahan prediksi keterlambatan adalah *plus* 2.21 menit atau minus 2.21 menit. Hal ini cukup baik karena rata – rata untuk andil yang di prediksi 5.57 menit dan standar deviasinya 6.22 menit. Dari scipy diambil library stats untuk menghitung *confidence*

interval. Untuk *confidence interval* 95% didapat kan keterlambatan berkisar diantara 3.2 menit sampai 5.52 menit.

Tabel 5 Contoh Hasil Prediksi

No	Andil	Andil Prediksi
1	25	19.8
2	-5	-4.3
3	-4	-3.4
4	-4	-5.3
5	1	1.1
6	2	5.8

3. Kesimpulan dan Saran

Berikut ini merupakan kesimpulan yang dihasilkan.

1. Pemodelan predictive analytics yang memiliki akurasi paling tinggi adalah dengan *ensemble learning* yang menggabungkan *decision tree regressor, support vector machine, random forest regressor dan gradient boosting.* Pecarian *hyperparameter* yang optimal untuk *random forest regressor* menggunakan GridSearchCV
2. Model prediksi memiliki akurasi 73.9% berdasarkan *cross validation* scoring yang dilakukan oleh sklearn. Model prediksi memiliki *mean absolute error* sebesar 2.21 menit dimana rata – rata untuk andil yang di prediksi 5.57 menit dan standar deviasinya 6.22 menit. Model prediksi untuk *confidence interval 95%* didapat kan keterlambatan berkisar diantara 3.2 menit sampai 5.52 menit
3. Untuk nomor kereta 395, yaitu jalur Cibatu Purwakarta, andil berkisar antara 10 menit – 50 menit untuk bulan February 2019 sampai bulan Maret 2019. Pada bulan Februari, terdapat 15 kasus keterlambatan sedangkan pada bulan Maret terdapat 17 kasus keterlambatan. Keterlambatan pada minggu 10 terjadi setiap hari sedangkan untuk minggu lainnya berkisar antara 1 hari sampai 4 hari. Keterlambatan pada umumnya terjadi pada hari jumat dan sabtu dibandingkan dengan hari lainnya (0=Senin). Keterlambatan paling sering terjadi pada lokasi Bandung diikuti oleh Rendeh, Sasaksaat, Andir, Cimekar, Padalarang, Sukatani, Cilame, dan Karangsari untuk keterlambatan yang lebih dari 10 kali. Untuk penyebab keterlambatan, terdapat PEMASANGAN TASPAT (PRASARANA), PENAMBAHAN HSD, TUNGGU PERSILANGAN, PEKERJAAN JEMBATAN yang memiliki frekuensi lebih dari 30 kali.
4. Untuk nomor kereta 396, yaitu jalur Purwakarta Cibatu, andil berkisar antara 10 menit – 50 menit untuk bulan Februari 2019 sampai bulan Maret 2019. Pada bulan Februari, terdapat 18 kasus keterlambatan sedangkan pada bulan Maret terdapat 16 kasus keterlambatan. Keterlambatan pada minggu 7 dan 11 terjadi 6 hari sedangkan untuk minggu lainnya berkisar antara 2 hari sampai 5 hari tiap minggunya. Keterlambatan pada umumnya terjadi pada hari Senin, tetapi keterlambatan untuk hari lainnya tidak begitu berbeda jauh, hanya pada haru Rabu saja memiliki keterlambatan hanya 2 kali (0=Senin). Keterlambatan paling sering terjadi pada lokasiPurwakarta, diikuti oleh Cicalengka, Bandung, Cilame, Cikudapateuh, Maswati, dan Nagreg, untuk keterlambatan yang lebih

dari 10 kali. Untuk penyebab keterlambatan, terdapat PEMASANGAN TASPAT (PRASARANA), TUNGGU PERSILANGAN, PEKERJAAN JEMBATAN yang memiliki frekuensi lebih dari 30 kali.

Berikut ini merupakan saran yang dapat dijadikan masukan untuk penelitian selanjutnya.

1. Menggunakan *neural network* atau *deeplearning* untuk pemodelan *machine learning*. Penggunaan neural network dan deep learning diperlukan apabila jumlah data sangat besar sehingga dapat mencakup lebih banyak jalur pada PT. KAI.
2. Menambahkan jumlah data, dari segi waktu dan dari segi jumlah rute untuk meningkatkan akurasi dan mendapatkan wawasan yang lebih baik mengenai permasalahan keterlambatan, terutama berdasarkan penyebab keterlambatan tersebut. Menambahkan pula jumlah data untuk setiap jenis penyebab agar akurasi untuk setiap penyebab keterlambatan dapat menjadi lebih baik lagi.

Daftar Pustaka

[1] Robert N, Kim H, "Predictions of *train* delays using *machine learning*", 2018

[2] Ramashish G, Biplav S, "Estimating *Train* Delays in a Large Rail Network Using Zero Shot Markov Model, 2018

[3] Pu W, Qing-peng Z,"*Train* delay analysis and prediction based on big data fusion", 2019

[4] William A R, Gary E, "The Team Data Science Process lifecycle", 2020, diambil dari https://docs.microsoft.com/en-us/azure/machine-*learning*/team-data-science-process/lifecycle pada tanggal 2 Maret 2020 Pukul 15.31

[5] T.Chai, R.R. Draxler, "Root mean square error (RMSE) or *mean absolute error* (MAE)? – Arguments against avoiding RMSE in the literature", 2014

[6] Aurelien G, "Hands On *Machine learning* With Scikit-Learn & Tensorflow", 2017

[7] Olsson NOE, Haugland H. Influencing factors on train punctuality—results from some Norwegian studies. *Transp Policy* 2004;**11**:387–397.