LAPORAN

DATA ANALYTICS COMPETITION

FIND IT! 2021

Tim Dengan Nama Tim



Disusun Oleh :

Alif Adwitiya Pratama

Bagja 9102 Kurniawan

Pratama Azmi Atmajaya

Bandung

2021

DAFTAR ISI

[DAFTAR GAMBAR 3](#_Toc72706091)

[DAFTAR TABEL 3](#_Toc72706092)

[1. Latar Belakang 3](#_Toc72706093)

[2. Tujuan 4](#_Toc72706094)

[3. Metode Analisis Data 4](#_Toc72706095)

[a. Regresi Linear 4](#_Toc72706096)

[b. Random Forest 5](#_Toc72706097)

[c. SVR ( Support Vector Regression ) 5](#_Toc72706098)

[d. KNN ( K-Nearest Neighbor ) 5](#_Toc72706099)

[e. MLP ( Multi Layer Perceptron ) 6](#_Toc72706100)

[4. Analisis 6](#_Toc72706101)

[A. EDA 6](#_Toc72706102)

[B. Preprocessing 9](#_Toc72706103)

[C. Feature engineering 11](#_Toc72706107)

[1. Kolom rating\_value\_0 & rating\_value\_1 12](#_Toc72706108)

[2. Kolom Weight 12](#_Toc72706109)

[3. Kolom dimension\_n 12](#_Toc72706110)

[4. Kolom lexile\_measure 13](#_Toc72706111)

[5. Kolom reading\_age 13](#_Toc72706112)

[6. Kolom grade\_level 14](#_Toc72706113)

[D. Modelling 14](#_Toc72706114)

[1. Semi Supervised 14](#_Toc72706115)

[2. Predicting test set ( Supervised Learning ) 16](#_Toc72706117)

[E. Evaluation 18](#_Toc72706119)

[1. Semi Supervised 18](#_Toc72706120)

[2. Predicting test Set 20](#_Toc72706122)

[5. Kesimpulan 21](#_Toc72706124)

[6. Daftar Pustaka 22](#_Toc72706125)

# DAFTAR GAMBAR

[Gambar 4.B.1 10](#_Toc72706209)

[Gambar 4.B.2 11](#_Toc72706210)

[Gambar 4.B.3 12](#_Toc72706211)

# DAFTAR TABEL

[Tabel 4.D.1.1 16](#_Toc72706297)

[Tabel 4.D.2.1 18](#_Toc72706299)

[Tabel 4.E.1.1 20](#_Toc72706302)

[Tabel 4.E.2.1 22](#_Toc72706304)

# Latar Belakang

Dewasa ini teknologi informasi telah merambat ke segala golongan mulai dari bayi sampai lanjut usia menikmati berbagai benefit dari teknologi informasi, baik di dunia kesehatan ,transportasi, pendidikan , industri dan masih banyak lagi. Tentu terdapat suatu aspek yang sangat krusial pada pengembangan kualitas hidup masyarakat suatu negara. Yaitu , pendidikan, pendidikan wajib ditempuh oleh setiap warga negara untuk mencapai cita cita negara tersebut. Salah satunya adalah dengan belajar , belajar dapat membuat sebuah negara memiliki SDM yang unggul pada dunia internasional yang siap untuk bersaing, untuk mewujudkan hal tersebut kebutuhan akan sebuah buku itu menjadi kebutuhan wajib bagi setiap pelajar di dunia, kebutuhan ini akan bertambah seiring naiknya pertumbuhan penduduk dan penambahan pelajar, tentu banyak sekali aspek aspek yang mempengaruhi minat seorang pelajar dalam membaca ataupun membeli buku, mulai dari tingkat kesulitan buku ,ekonomi, sampai dari minat dari pembaca. Selain itu, setiap industri yang menjual berbagai buku perlu mengetahui bagaimana memberikan harga berdasarkan karakteristik buku, sehingga hal ini akan memudahkan penerbit maupun reseller buku dalam melakukan penjualan kembali dengan harga yang sesuai dan saling untung baik untuk penerbit maupun penulis dan tidak lepas pula kepada para pelajar yang hendak membeli buku tersebut. Hal ini mendasari tujuan proposal ini dibuat, untuk mendukung hal ini , akan dibentuk suatu model machine learning yang dapat mengidentifikasi harga suatu buku berdasarkan karakteristik buku tersebut.

# Tujuan

Makalah ini bertujuan untuk mendeskripsikan teknik eksplorasi, pemrosesan, pembangunan model kepada pembaca dan menghasilkan suatu model yang dapat memprediksi harga buku berdasarkan karakteristik buku tersebut dengan baik dan sesuai supaya membantu meningkatkan minat pelajar dalam membaca buku.

# Metode Analisis Data

Pekerjaan pengolahan data dari awal sampai akhir menggunakan bahasa pemrograman python dengan library yang ada di dalamnya seperti pandas untuk mengolah *dataframe*, numpy untuk pengolahan array, seaborn & matplotlib untuk visualisasi data, scikit learn untuk pemodelan machine learning, dan tensorflow untuk pemodelan deep learning. Dengan model yang diuji coba pada dataset ini antara lain

## Regresi Linear

Merupakan metode statistika yang digunakan untuk memodelkan sebuah pengaruh satu atau beberapa variabel terhadap satu buah variabel yang disebut variabel respon dan variabel prediktor. Bila variabel prediktor berjumlah lebih dari satu digunakan analisis regresi linier berganda. Metode ini merupakan metode simple yang dapat menghasilkan prediksi yang baik jika data yang dimasukan cenderung sedikit dengan variansi yang tidak berbeda jauh. Namun, sebaliknya jika dimasukan memiliki variansi tinggi atau hubungan yang tidak linier maka model akan menjadi tidak akurat.

## Random Forest

Random forest (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi random forest dilakukan melalui penggabungan pohon (tree) dengan melakukan training pada sampel data yang dimiliki. Penggunaan pohon (tree) yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan random forest diambil berdasarkan hasil voting dari tree yang terbentuk. Pemenang dari tree yang terbentuk ditentukan dengan vote terbanyak. Pembangunan pohon (tree) pada random forest sampai dengan mencapai ukuran maksimum dari pohon data. Random forest memiliki kompleksitas waktu yang cukup besar namun dengan menggunakan random forest pada klasifikasi data maka, akan menghasilkan vote yang paling baik.

## SVR ( Support Vector Regression )

Support Vector Machine merupakan metode regresi, mempertahankan semua fitur utama yang menjadi karakteristiknya (margin maksimal). Support Vector Regression (SVR) menggunakan prinsip yang sama dengan SVM untuk klasifikasi, hanya saja dengan sedikit perbedaan kecil. Pertama-tama, karena keluaran adalah bilangan real, menjadikan model ini sangat sulit untuk memprediksi informasi yang ada dan memiliki kemungkinan tak terbatas. Dalam kasus regresi, margin toleransi (epsilon) ditetapkan mendekati SVM yang sudah diminta dari masalah. Namun, dilakukan untuk meminimalkan kesalahan, mengindividualisasikan hyperplane yang memaksimalkan margin, dengan membuat sebagian kesalahan dapat ditoleransi.

## KNN ( K-Nearest Neighbor )

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan data latih yang sudah terklasifikasikan sebelumya. Termasuk dalam supervised learning, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam KNN menggunakan euclidean distance. KNN merupakan salah satu algoritma (model) pembelajaran mesin yang bersifat nonparametrik. Yaitu model yang tidak mengasumsikan apa-apa mengenai distribusi instance di dalam dataset. Model nonparametrik biasanya lebih sulit diinterpretasikan, namun salah satu kelebihannya adalah class decision line yang dihasilkan model tersebut bisa jadi sangat fleksibel dan nonlinear sehingga dapat melakukan klasifikasi dengan tepat.

## MLP ( Multi Layer Perceptron )

model yang terinspirasi oleh bagaimana neuron dalam otak manusia bekerja. Tiap neuron pada otak manusia saling berhubungan dan informasi mengalir dari setiap neuron tersebut berubah untuk mempelajari sesuatu. Sama seperti itu sebuah MLP merupakan sekumpulan matriks yang memperbaiki nilainya setiap kali ada data baru masuk ke dalam input sehingga sebuah MLP dapat mempelajari pola dari data yang diberikan sehingga dapat dimanfaatkan untuk melakukan prediksi sesuai dengan pola yang ada pada data. MLP merupakan sebuah model yang cukup kompleks sehingga memakan banyak sekali resource untuk melatihnya serta waktu dan data yang banyak supaya hasil yang dihasilkan lebih maksimal. Namun, hal ini cukup sesuai dengan nilai akurasi yang dihasilkan akan lebih tinggi dibanding menggunakan model yang sebelumnya digunakan ( jika data yang digunakan dan arsitektur tepat )

# Analisis

## EDA

Exploratory data analysis digunakan untuk memahami data dengan melakukan hal tersebut akan didapat banyak sekali informasi guna menyelesaikan suatu masalah. Pada kasus dataset ini, hal yang pertama dilakukan adalah mengetahui secara general informasi informasi yang terdapat pada dataset. Seperti banyaknya row dan kolom, nan values, sebaran data, Hubungan antar kolom dan juga potensi pengembangan suatu kolom. Hasil daripada eksplorasi yang dilakukan adalah sebagai berikut.

1. Terdapat banyak sekali Kolom NaN values, sehingga hal ini perlu dilihat lebih dalam pada kolom kolom nya, yang didapatkan adalah NaN values ini bisa saja sebagai default values bahwa memang informasi tersebut tidak ada atau dapat direpresentasikan sebagai 0.
2. Terdapat beberapa kolom kategorikal yang tidak seimbang dan salah inputan, sebagai contoh kolom book format memiliki inputan yang tidak tepat, sehingga perlu dibenahi lebih lanjut.
3. Kolom price, dimana terdapat banyak NaN values, sehingga perlu dilakukan pendekatan semi supervised learning untuk menyelesaikan studi kasus ini
4. Sebaran data numerik tidak bersifat linear, sehingga penggunaan model statistik seperti regresi linear, elastic net, dan lasso tidak diperlukan.
5. Sebaran data bersifat non linear, sehingga dapat digunakan model model ensemble, SVR, maupun voting regressor dan Deep Learning.
6. Sebaran data bersifat non linear, sehingga dapat digunakan model model ensemble, SVR, maupun voting regressor dan DeepLearning.
7. Kolom dimension\_0,dimension\_1,dimension\_2 memiliki sebuah hubungan panjang, lebar,tinggi sehingga ke tiga kolom ini dapat direpresentasikan menjadi 1 kolom
8. BookFormat Didominasi dengan format yang tidak diketahui, namun untuk kasus 500 data yang kolom pricenya tidak NaN, didominasi dengan bookformat HardCover disusul oleh Paperback.
9. Kolom publisher\_date memiliki banyak sekali NaN values, sehingga tidak dapat diproses lebih jauh lagi, dengan begitu kolom ini akan dihapus pada saat modelling dan preprocessing nanti
10. Kolom publisher\_id,author\_id tidak diproses menjadi sebuah fitur lebih dalam, dikarenakan banyaknya nilai unik yang ada pada kolom tersebut, ditambah lagi kurangnya informasi terhadap kedua informasi diatas.
11. Kolom Description tidak diproses mengingat, membutuhkan threshold tersendiri dan labelling tersendiri dan juga penentuan dan pembentukan model, yang dapat membuat komputasi lebih jauh meningkat, selain itu terdapat beberapa row dengan kondisi NaN.
12. Kolom Lexile\_measure dapat diproses lebih lanjut, dikarenakan banyak informasi yang berada pada kolom ini, mulai dari kategorik lexilenya sampai dari pada nilai lexile nya sehingga dapat dijadikan suatu kolom tambahan nantinya.
13. Kolom grade\_level dapat diproses lebih lanjut, hal ini didasari dari pada kolom lexile\_measure yang didefinisikan, dengan adanya grade\_level, tentu dapat dijadikan sebuah basis terhadap buku yang berada pada suatu user. Sehingga, kolom ini dapat diproses lebih jauh lagi untuk menghasilkan kolom baru.
14. Kolom reading\_age dapat diproses lebih lanjut, hal ini didasari dari pada informasi lexile\_measure, untuk mendukung lexile\_measure ini sendiri diperlukan juga informasi terkait reading\_age, sehingga untuk reading age ini dapat diproses lebih dalam untuk menghasilkan suatu kolom baru.
15. Kolom rating\_value\_0 & 1

Kolom rating\_value\_0 & 1 adalah kolom dari rating sebuah buku pada website / tempat, hal ini penting mengingat harga buku juga dipengaruhi dari rating buku tersebut

1. Kolom genre\_n

Kolom ini tidak akan di hilangkan mengingat terdapat hubungan dengan genre\_n\_weight sehingga jika dihilangkan akan membuat model loss information.

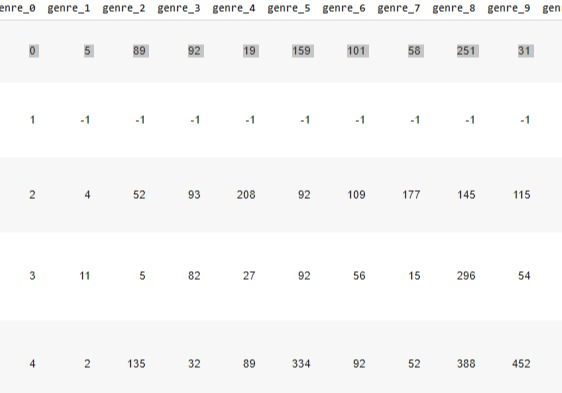
1. Pages

Untuk kolom pages, tidak akan dihilangkan mengingat pages adalah total halaman sebuah buku berdampak langsung terhadap harga dari buku tersebut.

1. Terdapat beberapa kolom yang berpotensi untuk memberikan insight baru atau kolom baru apabila diolah lebih dalam lagi, berdasarkan EDA, pengkodean setiap data kategorikal diperlukan mengingat model hanya menerima inputan berupa number
2. Kandidat Kolom yang dapat diproses lebih lanjut diantaranya adalah kolom lexile\_measure, grade\_level, reading\_age, rating\_value\_0, rating\_value\_1, weight, dan dimension\_n
3. Untuk buku diatas harga 150.000 didominasi oleh buku berformat Hardcover, hal ini wajar mengingat hardcover adalah buku cetaknya.
4. Terdapat beberapa pengecualian terhadap penggunaan one hot encoding pada beberapa kolom, seperti kolom genre\_1 sampai genre\_n, bookformat dan book edition, dikarenakan dapat menciptakan curse of dimensionality dan meningkatkan kompleksitas yang tinggi
5. Performansi terbaik akan diraih oleh model ensemble dikarenakan tingkat kompleksitas data yang tinggi.

## B. Preprocessing

Hal pertama yang dilakukan adalah dengan mengubah nilai yang kategorik menjadi angka agar model mudah untuk memproses data. Diantaranya ada kolom author\_id, publisher\_id, bookformat, bookedition, genre(0-9). Disini kita mulai mendaftarkan tiap nilai unik dari kolom tersebut lalu membuat sebuah kamus dari data unik tersebut. Setelah itu kamus diaplikasikan kedalam nilai kolom sehingga semua value di dalam list dapat berubah menjadi sebuah angka seperti gambar berikut



# Gambar 4.B.1

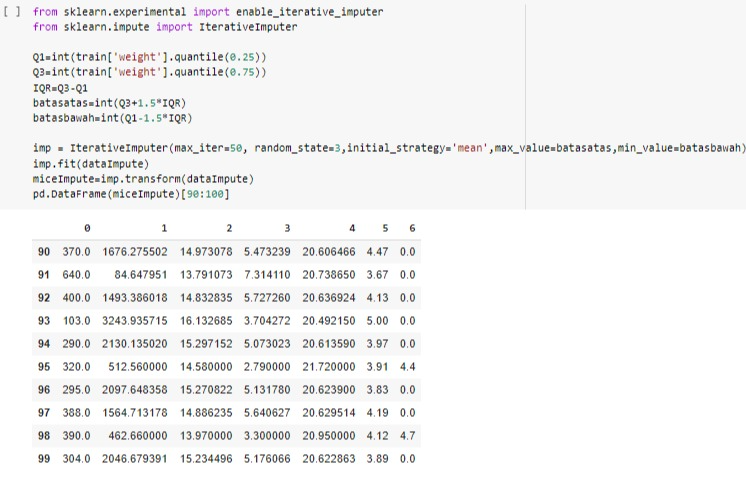
Dilanjutkan untuk memproses nilai kosong dari masing - masing kolom. Perlakuan pengisian null ini berbeda tiap kolomnya, untuk kolom rating\_value\_1 ditemukan bahwa setiap nilai null berpengaruh dengan kolom rating\_count\_1 apabila kolom ini bernilai ≤ 1 maka nilai dari rating\_value\_1 nya akan bernilai 0. Hal seperti ini ditemukan juga pada kolom genre\_weight apabila ada kekosongan pada kolom genre maka nilai genre weight nya akan mengikuti kosong karena itu kami mengisi kekosongan di weight dengan angka 0 sedangkan pada kolom genrenya telah otomatis terisi -1 karena kamus diatas otomatis mengubah nilai null menjadi angka -1. Null ini dapat diartikan bahwa buku memiliki ≤ 9 genre sehingga menyebabkan ada kolom genre yang kosong.

Untuk kolom seperti pages, weight, dimension(0-3) diisi menggunakan sebuah imputer dari library sklearn yang akan mengisi ke 3 kolom tersebut dengan bantuan kolom rating\_value\_0 dan rating\_value\_1 sebagai tambahan feature masukan



# Gambar 4.B.2

Yang selanjutnya akan menjadi seperti berikut setelah dilakukan pengisian



# Gambar 4.B.3

Imputer ini bekerja dengan mempelajari kolom yang tidak kosong dan berusaha mengisi kolom - kolom yang kosong sesuai dengan pola data sebelumnya. Setelah itu kita kembalikan ke dataframe utama kolom yang sudah terisi.

Sedangkan untuk kolom description, published\_date akan di drop karena tidak digunakan oleh model yang kami bangun. Namun, akan digantikan dengan feature buatan yang akan dibahas pada bab selanjutnya

## C. Feature engineering

Untuk feature engineering sendiri diambil beberapa kolom yaitu, kolom lexile\_measure, grade\_level, reading\_age, rating\_value\_0, rating\_value\_1, weight, dan dimension\_n. Seperti yang didapat pada EDA sebelumnya

### Kolom rating\_value\_0 & rating\_value\_1

Dimulai dari nilai rating\_value\_0 dan rating\_value\_1, akan dibentuk sebuah kolom kategorik, nilai kolom kategorik dibentuk dari pada range nilai rating\_value\_0 dan nilai rating\_value\_1. Yaitu tidak, sangat buruk, buruk, cukup, lumayan, baik, sangat baik. bentukan kategorikal diatas didapat dari pada pertimbangan nilai suatu buku pada sebuah website. Dengan melakukan representasi tersebut diharapkan model mendapati informasi yang lebih untuk memahami data.

### Kolom Weight

#### Mengkategorikan Weight

Dari kolom weight dapat ditambahkan kategorisasi dari berat weight yaitu dengan membaginya dalam 5 kelompok sesuai dengan aturan yang ditentukan, kelima kelompok tersebut adalah sangat ringan, ringan, sedang, berat, dan sangat berat. Perspektif yang kami gunakan adalah perspektif buku yang normal pada biasanya.

#### Weight\_per\_page

Setiap buku memiliki berat halamannya masing masing, pada kasus ini diasumsikan seluruh halaman memiliki berat kertas yang sama, maka dari itu dilakukan pengadaan fitur weight\_per\_page dengan memanfaatkan kolom weight dengan kolom pages.

### Kolom dimension\_n

Kolom dimension\_0, dimension\_1, dimension\_2 memiliki makna panjang, lebar,dan tinggi dari pada buku tersebut, sehingga apabila dikalikan didapatlah volume dari pada buku tersebut, sehingga dilakukan lah penghapusan kolom dimension\_n yang telah direpresentasikan dengan volume.

### Kolom lexile\_measure

#### Lexile\_codes

Pada kolom lexile\_measure dilakukan extract informasi pada kolom lexile\_measure, dikarenakan terdapat beberapa kategori pada lexile\_score itu sendiri seperti ad, hl,comic, dan lainnya. Dengan menggunakan regex hal tersebut dapat di extract, yang diharapkan mampu memaksimalkan kinerja model.

#### Lexile\_measureNumber

Kolom ini dihasilkan dengan melakukan preprocessing terlebih dahulu,dengan menghapus alphabet pada kolom lexile\_measure, sehingga hanya angkanya saja yang didapatkan. Hal ini bertujuan untuk mengetahui tingkatan buku tersebut, tentu hal tersebut dicapai dengan menggunakan regex.

#### Lexile\_categorical

Kolom ini dihasilkan dengan memberikan rules kolom lexile\_measureNumber sebagai berikut, hal ini didasarkan pada rules yang telah kami dapat pada website berikut <https://support.lexile.com/s/article/How-does-the-Lexile-scale-help-explain-proficiency-standards>, namun kami melakukan perubahan agar sesuai dan dapat diaplikasikan pada data.

### Kolom reading\_age

Sama seperti kolom lexile\_categorical di kolom reading\_age kita memperkaya fitur dengan cara membuat kategorisasi dari kolom reading\_age menjadi hanya beberapa kategori saja, dikarenakan kategori sebelumnya memiliki banyak redundansi, dan semakin banyak unik kategorikal pada sebuah kolom bukanlah suatu practice yang baik, maka dari itu dilakukan pengkonversian nilai unik tersebut menjadi beberapa kolom saja. Kategorisasi didasarkan pada tingkat pendidikan yang ditempuh oleh pelajar, terdapat TK, SD, dan others hal ini diharapkan dapat memberikan suatu informasi tambahan untuk model.

### Kolom grade\_level

Untuk kolom grade\_level memiliki behavior yang sama dengan kolom reading\_age sehingga redundansi ini diatasi dengan pembuatan beberapa kolom saja yaitu menjadi sebuah tingkatan baca dengan melihat perspektif seseorang yang memiliki kemampuan untuk membaca, yaitu very easy, easy, medium, dan Hard. Hal ini didasari dari kemampuan seseorang membaca suatu buku.

## D. Modelling

### Semi Supervised

Hal yang pertama kami lakukan setelah EDA, preprocessing, dan fitur engineering adalah filtering data yang tidak memiliki NaN Price, sehingga hanya didapat data data yang memiliki nilai price saja, hal ini dijadikan suatu basis pada penyelesaian studi kasus ini. Selanjutnya adalah melakukan scaling pada Fitur dan target variable, hal ini dilakukan guna menyamakan seluruh unit yang terdapat pada dataset. Setelah itu dilakukan lah splitting data, untuk validasi maupun training, hal ini guna untuk melihat performansi model random forest. Tentu kami melakukan modelling dengan menggunakan beberapa algoritma machine learning maupun deep learning yang ternyata hasilnya sebagai berikut :

|  |  |  |  |
| --- | --- | --- | --- |
| No | Model | MSE | RMSE |
| 1 | Linear Regression | 0.7838 | 0.8853 |
| 2 | Decision Tree | 0.0230 | 0.1517 |
| **3** | **Random Forest** | **0.0113** | **0.1065** |
| 4 | SVR | 0.0242 | 0.1556 |
| 5 | KNN | 0.0132 | 0.1150 |
| 6 | Voting Regressor | 0.0131 | 0.1148 |
| 7 | DeepLearning | 0.0135 | 0.1164 |

# Tabel 4.D.1.1

Pada tabel diatas didapat informasi MSE dan RMSE dimana model Random Forest adalah model yang memiliki performansi terbaik dalam melakukan fitting model. Model Linear Regression memiliki performansi terburuk, sesuai dari pada EDA yang telah dilakukan, dikarenakan data tidak berbentuk linear, sedangkan untuk KNN, dikarenakan model ini bersifat lazy sehingga, ia sangat bergantung pada sebaran data, apabila melihat kepada EDA didapat bahwasanya data yang dihasilkan memiliki sebaran data yang cukup baik untuk beberapa kolom, sehingga model KNN ini dapat melakukan fitting yang well terhadap data training. Selanjutnya adalah Decision Tree, model ini fitting data dengan baik, namun tentu untuk generalisasi decision tree wajib dipertanyakan dikarenakan model ini sangat rentan terhadap overfitting pada data. Untuk model Voting Regressor ini, ternyata memiliki performansi yang mirip dengan KNN dengan selisih nilainya 0.0001 tentu hal ini sebaiknya digunakan KNN mengingat voting regressor menggunakan komputasi lebih banyak, dimulai dari banyaknya model ditambah kasus regression yang membuat model ini bergantung kepada luaran masing masing model apabila terdapat outlier pada luarannya,maka tentu hasilnya tidak dapat diandalkan. Sedangkan untuk model MLP didapat bahwasanya model ini fitting dataset dengan baik,dimana ia hampir setara dengan KNN dan voting regressor, namun hal ini tentu tidak dapat diandalkan mengingat komputasi MLP lebih berat dibandingkan voting regressor maupun KNN, sehingga ada baiknya penggunaan model dialihkan menjadi model machine learning saja. Sedangkan untuk SVR, ternyata memiliki fitting yang kurang baik dibandingkan model KNN dan Voting regressor, hal yang mungkin mendasari luaran tersebut adalah sebaran data yang tidak mampu di tangkap oleh model SVR, sehingga menyebabkan perfomansi yang kurang baik. Random Forest, memiliki performansi terbaik diantara model lainnya, ini dikarenakan model random forest memiliki banyak estimator dan juga pembatasan pada kedalaman tree, sehingga akan menghindari model dari overfitting. Ditambah model ini didukung oleh banyaknya kategorikal dalam dataset, sehingga membuatnya mudah untuk fitting data.

### Predicting test set ( Supervised Learning )

Dalam melakukan prediksi pada test set diperlukan preprocessing data terlebih dahulu, seperti melakukan pengkodean pada kolom kategorikal, mengisi NaN values, melakukan feature engineering, dan melakukan beberapa hal lainnya agar sesuai dengan inputan model yang telah dibentuk, setelah itu dilakukan lah predict pada seluruh test set yang menghasilkan nilai sebagai berikut:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Model | MSE | RMSE |
| 1 | Linear Regression | 2305115.040 | 1518.260 |
| 2 | Decision Tree | 0.07184 | 0.2680 |
| 3 | SVR | 0.0297 | 0.1724 |
| 4 | KNN | 0.0408 | 0.2021 |
| 5 | Random Forest | 0.0292 | 0.1710 |
| 5 | Voting Regressor | 0.0377 | 0.1942 |
| 6 | Deep Learning | 0.0393 | 0.1983 |

# Tabel 4.D.2.1

Pada tabel diatas didapat informasi MSE dan RMSE dimana model Random Forest adalah model yang memiliki performansi terbaik dalam melakukan fitting model. Model Linear Regression memiliki performansi terburuk, sesuai dari pada EDA yang telah dilakukan, dikarenakan data tidak berbentuk linear, dan benar saja pada saat melakukan testing ternyata memberikan MSE yang sangat besar, hal ini mengindikasikan model linear regression sangat buruk dalam predicting new dataset, sehingga dapat disimpulkan tidak dapat generalisasi dengan baik. sedangkan untuk KNN, dikarenakan model ini bersifat lazy sehingga, ia sangat bergantung pada sebaran data, apabila melihat kepada EDA didapat bahwasanya data yang dihasilkan memiliki sebaran data yang cukup baik untuk beberapa kolom, dan pada saat melakukan predicting ternyata luarannya dapat diandalkan, hal ini menandakan bahwasanya sebaran data dengan parameter yang di pass pada training ternyata mampu generalisasi dengan baik. Selanjutnya adalah Decision Tree, model ini menghasilkan predicting yang cukup baik, namun kembali lagi, dikarenakan decision tree sangat rentan terhadap overfitting pada data. Walaupun begitu, predicting hasilnya cukup memuaskan. Untuk model Voting Regressor ini, ternyata memiliki performansi yang lebih baik dari pada dengan KNN namun sebaiknya mempertimbangkan kembali penggunaan model ini mengingat voting regressor menggunakan komputasi lebih banyak, dimulai dari banyaknya model ditambah kasus regression yang membuat model ini bergantung kepada luaran masing masing model apabila terdapat outlier pada luarannya,maka tentu hasilnya tidak dapat diandalkan. Sedangkan untuk model MLP didapat bahwasanya model ini predicting test set cukup baik, namun jika dibandingkan dengan model Machine learning seperti SVR, Random Forest, dan voting regressor masih kurang. Tentu hal ini tidak dapat diandalkan mengingat komputasi MLP lebih berat dibandingkan voting regressor maupun random forest, dan SVR, sehingga ada baiknya penggunaan model dialihkan menjadi model machine learning saja. Sedangkan untuk SVR, ternyata memiliki hasi predicting yang baik dibandingkan, hal yang mungkin mendasari luaran tersebut adalah sebaran datanya mampu ditangkap oleh model SVR, sehingga menyebabkan perfomansi yang cukup baik. Random Forest, memiliki performansi terbaik diantara model lainnya, ini dikarenakan model random forest memiliki banyak estimator dan juga pembatasan pada kedalaman tree, sehingga akan menghindari model dari overfitting. Ditambah model ini didukung oleh banyaknya kategorikal dalam dataset, sehingga membuatnya mudah untuk predicting data.

## E. Evaluation

Untuk melakukan evaluasi terhadap model model yang didapat kita melakukan beberapa perhitungan metrics sebagai berikut :

### Semi Supervised

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | Model | Mean | Median | STD | MAE | MSE | RMSE |
| 1. | Linear Regression | 0.1323 | 0.2262 | 0.8946 | 0.1703 | 0.7838 | 0.8853 |
| 2. | KNN | 0.2085 | 0.2188 | 0.0835 | 0.0782 | 0.0132 | 0.1150 |
| 3. | Decision Tree | 0.2132 | 0.1796 | 0.1274 | 0.1035 | 0.0230 | 0.1517 |
| **4.** | **Random Forest** | **0.2154** | **0.2268** | **0.0783** | **0.0699** | **0.0113** | **0.1065** |
| 5. | Voting Regressor | 0.2111 | 0.2328 | 0.0937 | 0.0760 | 0.0131 | 0.1148 |
| 6. | DeepLearning | 0.1974 | 0.2032 | 0.0816 | 0.0768 | 0.0135 | 0.1164 |

# Tabel 4.E.1.1

Kasus Linear Regression, apabila melihat dari pada rata rata, mediannya maupun standar deviasinya, dapat kita tarik kesimpulan bahwasanya hasil sebarannya tidak berdistribusi normal dan tidak berkonsentrasi pada titik meannya sehingga hal ini mempengaruhi nilai loss functionnya yaitu *Mean Absolut Error*, *Mean Squared Error*, sampai *Root mean squared error* nya akan sangat jelek, dan benar saja, dapat dilihat dari pada luarannya.

Kasus KNN, apabila melihat dari pada mean,median, dan std nya dapat ditarik sebuah kesimpulan bahwasanya luaran prediksinya berkonsentrasi pada titik mean, sehingga sebarannya banyak mendekati mean, dan tentu hal ini juga mempengaruhi fitting data yang baik, terlihat bahwasanya MAE, MSE, dan RMSE nya juga mendekati 0, yang berarti hasil fitting datanya dapat mereduksi nilai lossnya.

Kasus Decision Tree, untuk kasus ini jika kita melihat dari pada mean, median, dan std nya sebaran datanya tidak terlalu berkonsentrasi pada titik mean sehingga hal ini akan membuat MAE, MSE, dan RMSE naik.

Kasus Random Forest, apabila melihat dari sebaran data nya, dimulai dari mean, median, model ini menghasilkan luaran yang paling kecil dibandingkan model lainnya ini mengindikasikan bahwasanya model yang dihasilkan dapat menyesuaikan dengan baik. Dan hal ini dibuktikan daripada MAE, MSE, dan RMSE yang paling kecil diantara model lainnya.

Kasus Voting regressor, dari segi sebaran datanya sendiri voting regressor menghasilkan model yang hampir menyimbangi model random forest, terlihat bahwasanya sebarannya mirip dengan random forest, namun pada kasus ini random forest lebih baik dibandingkan model voting regressor, hal ini dapat disebabkan karena model model yang terdapat pada voting regressor itu memiliki luaran yang kurang baik, dan mempengaruhi nilai akhir dari pada hasil regresi model voting ini.

Kasus Deep Learning, dari segi sebaran data nya dapat dilihat bahwasanya hal unik dari luarannya adalah rata ratanya cukup tinggi dibandingkan model lainnya namun memiliki std yang sangat rendah, ini mengindikasikan bahwasanya sebaran data sangat berkonsentrasi pada mean, dan hal ini tentu berefek pada luaran yang dihasilkan , namun untuk kompleksitas model yang tinggi dan performansi yang masih dibawah daripada random forest. Tentu model ini belum dapat menggantikan Random Forest.

### Predicting test Set

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | Model | Mean | Median | STD | MAE | MSE | RMSE |
| 1. | Linear Regression | 1220.10 | 1048.89 | 903.81 | 1219.99 | 2305 | 1518 |
| 2. | KNN | 0.2941 | 0.2957 | 0.0042 | 0.1914 | 0.0408 | 0.2021 |
| 3. | Decision Tree | 0.2692 | 0.2303 | 0.1502 | 0.1808 | 0.0554 | 0.2354 |
| **4.** | **Random Forest** | **0.2590** | **0.2573** | **0.0372** | **0.1579** | **0.0292** | **0.1710** |
| 5. | Voting Regressor | 0.2772 | 0.2680 | 0.0461 | 0.1764 | 0.0370 | 0.1923 |
| 6. | DeepLearning | 0.2910 | 0.2910 | 0.0 | 0.1879 | 0.0394 | 0.1985 |

# Tabel 4.E.2.1

Dari keseluruhan informasi yang disajikan pada tabel didapatkan bahwasanya model yang dapat generalisasi yang lebih baik daripada model lainnya adalah model Random Forest , sesuai dari pada hasil EDA yang telah dilakukan. Tentu hal ini telah terlihat jelas dari pada karakteristik dataset , dan keberhasilan model semi supervised . Sehingga model dapat menghasilkan RMSE yang baik.

# Kesimpulan

Setelah dilakukan in depth analysis melalui EDA, preprocessing , feature engineering , modelling dan evaluation didapatlah sebuah kesimpulan, yaitu , melalui studi kasus regresi buku ini didapatkan bahwasanya informasi informasi esensial dan mendukung suatu target variable akan menghasilkan luaran yang mendukung pula. Dari hal studi kasus ini , seluruh variabel dan sebaran data perlu untuk dipahami lebih dalam lagi sehingga memunculkan informasi informasi tersembunyi pada dataset, yang tentunya akan membuat modelling lebih baik lagi. Selain itu, dilakukannya benchmarking model, agar pembaca dapat mengetahui performansi masing - masing model dengan data yang dilakukan secara jelas sehingga akan memperjelas kembali Penggunaan metrics metrics yang telah didefinisikan adalah dasar dalam melakukan evaluasi model yang mana hasilnya ternyata baik. Hasil dari pada proposal ini tentu dapat dikembangkan dan diaplikasikan pada studi kasus lain . Selain itu diharapkan hasil dari proposal dan model ini dapat bermanfaat dan meningkatkan minat baca . Perlu adanya penambahan dan penkoreksian dikarenakan masih banyak kekurangan pada proposal ini.

# Daftar Pustaka

Herawati, S. and Usman, M. (2003) ‘Analisis Beberapa Model Regresi Linear’, *jurnal matematika, aplikasi dan pembelajarannya*, 2(universitas negeri jakarta), p. 298.

*How does the Lexile scale help explain proficiency standards?* (no date). Available at: https://support.lexile.com/s/article/How-does-the-Lexile-scale-help-explain-proficiency-standards (Accessed: 23 May 2021).

INFORMATIKALOGI (2017) *Algoritma K-Nearest Neighbor (K-NN) | INFORMATIKALOGI*. Available at: https://informatikalogi.com/algoritma-k-nn-k-nearest-neighbor/ (Accessed: 23 May 2021).

jason, B. (2020) *Crash Course On Multi-Layer Perceptron Neural Networks*. Available at: https://machinelearningmastery.com/neural-networks-crash-course/ (Accessed: 23 May 2021).

Ma, J., Theiler, J. and Perkins, S. (2003) ‘Accurate On-line Support Vector Regression’, *Neural Computation*, 15(11), pp. 2683–2703. doi: 10.1162/089976603322385117.

Sharp, tom (2020) *An Introduction to Support Vector Regression (SVR) | by Tom Sharp ? | Towards Data Science*. Available at: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2 (Accessed: 23 May 2021).

Yanuar, A. (2018) *Random Forest – Universitas Gadjah Mada Menara Ilmu Machine Learning*. Available at: https://machinelearning.mipa.ugm.ac.id/2018/07/28/random-forest/ (Accessed: 23 May 2021).