Statistical Analysis on Factors Influencing Life Expectancy

Dina Lestari[1, a)], Dwi Ariyanti Nur Latifah[2, b)], Rahayu Isnaini[3, c)], Richwani Neysa Febriana[4, d)], Tara Dwipa Ardhatu Timur[5, e)], Tarisa Putri Cahyani[6, f)]

[1,2,3,4,5,6] *Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada, Indonesia*

[a)] Corresponding author: dinalestari057@mail.ugm.ac.id
[b]dwiariyanti02@mail.ugm.ac.id
[c)]rahayu.isnaini@mail.ugm.ac.id
[d)]richwani.neysa@mail.ugm.ac.id
[e)]taradwipa02@mail.ugm.ac.id
[f)]tarisa.putri.cahyani@mail.ugm.ac.id

Abstrak: Kondisi kesehatan sebuah negara dapat dilihat melalui angka harapan hidupnya atau *life expectancy*. Melalui model *supervised machine learning*, penelitian ini dilakukan untuk memprediksi angka harapan hidup di 193 negara di dunia untuk membandingkan kondisi kesehatan masyarakat negara maju dan berkembang. Dataset yang digunakan diambil dari situs Kaggle yang terdiri dari 22 variabel. Dalam proses perhitungan model prediksi digunakan metode analisis regresi dengan algoritma decision tree dan random forest. Model prediksi terbaik dipilih berdasarkan nilai mean squared error (MSE) dan mean absolute error (MAE) yang paling kecil. Dari prediksi yang telah dilakukan, didapatkan model regresi yang paling baik untuk dataset ini adalah model dengan algoritma random forest (susunan *hyperparameter* dengan metode *grid search*). Diharapkan hasil prediksi ini dapat digunakan untuk evaluasi pemerintah terkait sehingga dapat dilakukan perbaikan kondisi kesehatan masyarakat kedepannya.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kesehatan merupakan salah satu faktor penting yang mempengaruhi kesejahteraan suatu negara. Salah satu variabel yang dapat mengukur kesejahteraan suatu negara dapat dilihat dari angka harapan hidup atau *life expectancy*. Angka harapan hidup (AHH) adalah alat yang digunakan pemerintah untuk mengukur kinerja pemerintah dalam meningkatkan kesejahteraan dan kesehatan penduduk. Selain itu, AHH juga mengindikasikan derajat kesehatan masyarakat dan mencerminkan tingkat keberhasilan pembangunan bidang kesehatan. Semakin tinggi AHH maka derajat kesehatan masyarakat semakin baik dan hal ini didukung oleh keberhasilan dalam pembangunan bidang kesehatan. Sebaliknya, pembangunan bidang kesehatan yang kurang berhasil berdampak pada rendahnya derajat kesehatan masyarakat sehingga AHH rendah (Anggraini et al., 2013).

Penelitian mengenai angka harapan hidup sangat diperlukan karena indikator ini merupakan bentuk penilaian dari pembangunan ekonomi dan kesehatan yang telah dilakukan oleh suatu negara. Di beberapa negara yang tingkat kesehatannya lebih baik, setiap individu memiliki rata-rata hidup lebih lama yang berarti dari segi ekonomis mempunyai peluang untuk memperoleh pendapatan lebih tinggi. Oleh karena itu, pada penelitian ini akan dilakukan prediksi angka harapan hidup untuk 193 negara berdasarkan 22 variabel yang terdiri dari berbagai faktor, seperti faktor kesehatan, besarnya pendapatan, dan IPM.

Melalui penerapan *data mining*, prediksi yang dilakukan akan menggunakan bantuan dari *machine learning*. Dari berbagai macam metode yang terdapat pada *machine learning*, dipilih metode analisis regresi yang menggunakan algoritma *decision tree* dan *random forest*. Kedua algoritma tersebut dipilih karena akan dilakukan perbandingan nilai MSE dan MAE dari kedua algoritma tersebut dalam menentukan prediksi angka harapan hidup. Metode prediksi terbaik dipilih berdasarkan hasil prediksi yang memiliki nilai MSE dan MAE paling rendah.

Dari latar belakang yang telah dijelaskan di atas, maka akan dilakukan prediksi berdasarkan variabel-variabel yang ada pada dataset angka harapan hidup yang didapatkan dari WHO dan United Nations dengan metode analisis regresi algoritma *decision tree* dan *random forest*.

1.2 Tujuan dan Manfaat

Tujuan dari analisis ini adalah untuk memprediksi angka harapan hidup dari 193 negara di dunia berdasarkan data dari 22 variabel meliputi faktor kesehatan, ekonomi, dan sosial dengan menggunakan analisis regresi. Dua metode yang digunakan adalah Decision Tree dan Random Forest. Dari hasil analisis ini, diharapkan dapat digunakan pemerintah terkait untuk memajukan sektor kesehatan dan sosial sehingga dapat terwujud masyarakat yang sejahtera.

BAB II

LANDASAN TEORI

2.1 Regresi

Analisis regresi merupakan salah satu jenis *supervised learning* pada *machine learning* yang memprediksi nilai suatu variabel pada kumpulan data untuk memungkinkan analisis yang lebih akurat. Analisis regresi memiliki variabel respon berupa numerik, misalnya seperti memprediksi usia seseorang dari karakteristik orang tersebut

Evaluasi Model Regresi

Untuk mengevaluasi seberapa baik model yang dihasilkan oleh algoritma regresi, kita dapat mengevaluasinya menggunakan beberapa formula berikut, dimana model regresi yang baik merupakan model regresi yang memiliki error paling rendah.

1. Mean Absolute Error (MAE)
   MAE mengukur error pada skala yang sama dengan skala pada data karena hanya bergantung pada ei . Ukuran ini tidak dapat digunakan untuk membandingkan data dengan unit yang berbeda. Selain itu, MAE juga robust terhadap outlier

$$MAE = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{n}$$

dengan i = observasi ke-i, n adalah banyaknya data, yi = nilai observasi yang sebenarnya, dan yˆi = estimasi nilai observasi.

2. Root Mean Squared Error (RMSE)
   RMSE merupakan ukuran yang paling sering digunakan dalam mengevaluasi seberapa baik model regresi yang dihasilkan. RMSE juga mengukur error pada skala yang sama dengan skala pada data dan tidak cocok digunakan untuk membandingkan data dengan unit yang berbeda. RMSE lebih sensitif dibandingkan MAE karena pada RMSE, error akan dikuadratkan terlebih dahulu walaupun pada akhirnya juga akan diakarkan. Hal ini tentu saja baik karena RMSE mampu memberikan pinalti yang lebih tinggi semakin bertambahnya nilai error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

   dengan i = observasi ke-i, n = banyaknya data, yi = nilai observasi yang sebenarnya, dan yˆi = estimasi nilai observasi.

3. Mean Absolute Percentage Error (MAPE)
   MAPE merupakan ukuran yang unit-free sehingga umumnya digunakan untuk membandingkan peramalan dari data yang berbeda beda. Namun, MAPE akan menghasilkan nilai tak terbatas atau tidak terdefinisi ketika nilai actual suatu observasi adalah nol atau mendekati nol. MAPE juga asimetris dan memberikan pinalti yang lebih berat pada kesalahan negatif (ketika prediksi lebih tinggi dari yang sebenarnya) daripada kesalahan positif.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{100(y_i - \hat{y}_i)}{y_i} \right|$$

   dengan i = observasi ke-i, n = banyaknya data, yi = nilai observasi yang sebenarnya, dan yˆi = estimasi nilai observasi

2.2 *Decision Tree*

Pohon (tree) merupakan struktur data yang terdiri dari simpul (node) dan rusuk (edge). Simpul yang terdapat pada pohon dibedakan menjadi tiga bagian, yaitu simpul akar (root node), simpul percabangan/internal (branch/internal node), dan simpul daun (leaf node). Proses decision tree adalah mengubah data dalam tabel menjadi model pohon, kemudian mengubah pohon menjadi sebuah rule, lalu menyederhanakan rule. Tujuan dari metode ini adalah mem-breakdown proses pengambilan keputusan yang kompleks menjadi lebih sederhana sehingga keputusan yang diperoleh lebih akurat sebagai solusi dari suatu permasalahan. Decision

tree juga dapat digunakan untuk mengeksplorasi data, menemukan hubungan yang sebelumnya tidak diketahui antara variabel dependen dan independen

2.3 *Random Forest*

Random Forest adalah suatu metode yang terdiri dari sekumpulan pohon terstruktur yang masing-masing melemparkan unit suara untuk kelas dan hasil yang diperoleh berdasarkan keputusan terbanyak. Teknik dasar yang digunakan *Random Forest* adalah *Decision Tree.* Dengan kata lain, Random Forest adalah sekumpulan pohon keputusan (*Decision Tree*) yang digunakan untuk klasifikasi dan prediksi suatu data dengan memasukkan input ke bagian akar yang berada diatas kemudian turun ke bagian daun yang berada dibawah (Haristu, 2019). Pohon yang digunakan dalam metode ini bisa mencapai ratusan dengan cara penanaman yang sama bagi setiap pohonnya. *Random Forest* menggunakan strategi ensemble bagging yang dapat mengatasi masalah *overfitting* yang terjadi jika data train yang dimiliki berukuran kecil (Samudra, 2019). Hasil dari analisis *Random Forest* untuk klasifikasi merupakan modus dari setiap pohon dari hutan yang dibangun, sedangkan hasil prediksi diperoleh dari nilai rata-rata setiap pohon (Lingga P, 2017).

2.4 *Hyperparameter Grid Search*

Hyperparameter digunakan untuk mengatur berbagai macam aspek dalam machine learning yang sangat berpengaruh pada performa dan model yang dihasilkan. Pencarian hyperparameter dilakukan secara manual atau dengan menguji kumpulan *hyperparamete*r pada parameter yang ditentukan sebelumnya (Claesen & De Moor, 2015).

Salah satu metode hyperparameter yang dapat diaplikasikan adalah *grid search*. *Grid search* merupakan metode alternatif yang digunakan untuk menemukan parameter terbaik dalam suatu model, sehingga metode yang digunakan secara akurat memprediksi data yang digunakan. *Grid search* dikategorikan sebagai metode yang teliti, karena dalam menentukan parameter terbaik dilakukan eksplorasi masing masing parameter dengan mengatur jenis nilai prediksi terlebih dahulu. Kemudian metode tersebut akan menampilkan skor untuk masing-masing nilai parameter. *Grid Search* dapat diterapkan secara maksimum, apabila batas atas dan batas bawah dari masing masing parameter diketahui (Ramadhan et al., 2017)

BAB III

METODE PENELITIAN

3.1 Tentang Data

Jenis data yang digunakan adalah data sekunder. Data tersebut diperoleh dari situs web penyimpanan data WHO pada faktor kesehatan dan data ekonomi terkait dikumpulkan dari situs web Perserikatan Bangsa-Bangsa. File data individual telah digabungkan menjadi satu kumpulan data. Data ini dikumpulkan dalam 15 tahun yakni dari 2000 hingga 2015 pada 193 negara.

Data yang digunakan terdiri dari 2938 observasi dan 22 variabel. Variabel-variabel yang digunakan adalah sebagai berikut:

- *Country* : Negara
- *Year* : Tahun

- *Status* : Status negara meliputi *developing* atau *developed*
- *Life expectancy* : Angka Harapan Hidup
- *Adult Mortality* : Angka Kematian Dewasa dari kedua jenis kelamin (probabilitas kematian antara 15 dan 60 tahun per 1000 penduduk)
- *Infant deaths* : Jumlah Kematian Bayi per 1000 penduduk
- *Alcohol* : Alkohol, tercatat konsumsi per kapita (15+) (dalam liter alkohol murni)
- *Percentage Expenditure* : Pengeluaran untuk kesehatan sebagai persentase dari Produk Domestik Bruto per kapita (%)
- Hepatitis B : Cakupan imunisasi Hepatitis B (HepB) pada anak usia 1 tahun (%)
- *Measles* : jumlah kasus campak yang dilaporkan per 1000 penduduk
- *BMI* : Rata-rata Indeks Massa Tubuh dari seluruh populasi
- *Under-five deaths* : Jumlah kematian balita per 1000 penduduk
- *Polio* : Cakupan imunisasi Polio (Pol3) pada anak usia 1 tahun (%)
- *Total expenditure* : Pengeluaran pemerintah umum untuk kesehatan sebagai persentase dari total pengeluaran pemerintah (%)
- *Diphtheria* : Cakupan imunisasi difteri tetanus toksoid dan pertusis (DTP3) pada anak usia 1 tahun (%)
- *HIV/AIDS* : Kematian per 1.000 kelahiran hidup HIV/AIDS (0-4 tahun)
- *GDP* : Produk Domestik Bruto per kapita (dalam USD)
- Population : Populasi negara
- *Thinness 1-19 years* : Prevalensi ketipisan pada anak-anak dan remaja untuk usia 10 sampai 19 tahun (%)
- *Thinness 5-9 years* : Prevalensi ketipisan pada anak-anak untuk Usia 5 sampai 9 (%)
- *Income composition of resources* : Indeks Pembangunan Manusia dalam hal komposisi pendapatan sumber daya (indeks berkisar dari 0 hingga 1)
- *Schooling* : Jumlah tahun Sekolah (tahun)

Variabel yang digunakan dalam analisis ini ialah variabel dependen dan variabel independen. Variabel dependen yang digunakan adalah *Life expectancy* dan variabel independennya adalah *Country*, *Year*, *Status, Adult Mortality*, *infant deaths*, *Alcohol*, p*ercentage expenditure*, Hepatitis B, *Measles*, BMI, *under-five deaths*, Polio, *Total expenditure*, *Diphtheria*, HIV/AIDS, GDP, *Population*, *thinness 1-19 years*, t*hinness 5-9 years*, *Income composition of resources, Schooling*

3.2 Metode Analisis

Berdasarkan dataset *life expectancy* yang diperoleh ingin diprediksi *life expectancy* berdasarkan variabel independen yang ada, maka digunakan analisis regresi dengan menggunakan algoritma *decision tree* dan *random forest*. Dalam analisis ini, digunakan kedua algoritma tersebut karena ingin dibandingkan algoritma mana yang memberikan hasil yang lebih baik.

BAB IV

HASIL DAN PEMBAHASAN

4.1 *Data Importing*

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259210	33736494.0	17.2	17.3	0.479	10.1
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696514	327582.0	17.5	17.5	0.476	10.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744976	31731688.0	17.7	17.7	0.470	9.9
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959000	3696958.0	17.9	18.0	0.463	9.8
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537231	2978599.0	18.2	18.2	0.454	9.5

Tabel 1. Tabel dataset awal

Terlebih dahulu, dilakukan input dataset mengenai prediksi *life expectancy* berdasarkan variabel independen yang ada. Dilakukan pemanggilan lima data pertama dari dataset untuk mengecek kesesuaian data yang terinput dengan dataset yang dimiliki. Berdasarkan **Tabel 1** terlihat bahwa data yang terinput sudah sesuai dengan dataset.

4.2 *Data Preprocessing*

```
Country                           object
Year                               int64
Status                            object
Life expectancy                  float64
Adult Mortality                  float64
infant deaths                      int64
Alcohol                          float64
percentage expenditure           float64
Hepatitis B                      float64
Measles                            int64
 BMI                             float64
under-five deaths                  int64
Polio                            float64
Total expenditure                float64
Diphtheria                       float64
 HIV/AIDS                        float64
GDP                              float64
Population                       float64
 thinness  1-19 years            float64
 thinness 5-9 years              float64
Income composition of resources  float64
Schooling                        float64
dtype: object
```

Gambar 1. Deskripsi data

Sebelum dilakukan analisis data perlu dilakukan *preprocessing* terlebih dahulu. *Preprocessing* yang dimaksud disini adalah tahap mendasar dalam penambangan data untuk meningkatkan efisiensi data karena tahap ini data secara langsung memengaruhi hasil algoritma analitik apa pun. Pada tahapan awal dilakukan pengecekan tipe data dari masing-masing variabel. Berdasarkan **Gambar 1** didapatkan sebagian besar dataset dalam penelitian ini memiliki tipe

data numerik desimal (*float64*), selain itu juga didapatkan tipe data bilangan bulat (*int64*) dan tipe data unik (*object*).

```
Country                            0
Year                               0
Status                             0
Life expectancy                   10
Adult Mortality                   10
infant deaths                      0
Alcohol                          194
percentage expenditure             0
Hepatitis B                      553
Measles                            0
 BMI                              34
under-five deaths                  0
Polio                             19
Total expenditure                226
Diphtheria                        19
 HIV/AIDS                          0
GDP                              448
Population                       652
 thinness  1-19 years             34
 thinness 5-9 years               34
Income composition of resources  167
Schooling                        163
dtype: int64
```

Gambar 2. Pengecekan *missing value* pada data

Selanjutnya pada tahapan *preprocessing* adalah melakukan pengecekan mengenai ada atau tidaknya *missing value* pada dataset. Berdasarkan **Gambar 2** terlihat bahwa ditemukan banyak missing value pada variabel *Life expectancy, Adult mortality, Alcohol, Hepatitis B, BMI, Polio, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources,* dan *Schooling*. Oleh karena ditemukan banyaknya *missing value*, maka perlu dilakukan penanganan. Dalam menangani banyaknya *missing value* dapat digunakan imputasi atau pengisian pada nilai yang hilang dengan nilai mean karena data yang dimiliki merupakan data kontinu.

```
Country                          0
Year                             0
Status                           0
Life expectancy                  0
Adult Mortality                  0
infant deaths                    0
Alcohol                          0
percentage expenditure           0
Hepatitis B                      0
Measles                          0
 BMI                             0
under-five deaths                0
Polio                            0
Total expenditure                0
Diphtheria                       0
 HIV/AIDS                        0
GDP                              0
Population                       0
 thinness  1-19 years            0
 thinness 5-9 years              0
Income composition of resources  0
Schooling                        0
dtype: int64
```

Gambar 3. Penanganan missing value pada dataset

Setelah dilakukan penanganan dengan imputasi atau pengisian pada nilai yang hilang didapatkan secara keseluruhan **Gambar 3** pada masing masing variabel yaitu nol atau 0 yang berarti tidak ditemukan adanya *missing value* pada masing-masing variabel. Sehingga selanjutnya dapat dilakukan pembentukan variabel dummy pada dataset dengan menggunakan fungsi *pd.get_dummies*. Dengan menggunakan fungsi tersebut didapatkan variabel *dummy* termasuk untuk *reference category*. Sehingga perlu dilakukan pembuangan variabel *dummy* yang merupakan *reference category* untuk mencegah terjadinya multikolinearitas sempurna.

	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	...	Year_2007	Year_2008	Year_2009	Year_2010	Year_2011	Year_2012	Year_2013	Year_2014	Year_2015	Status_Developed
0	65.0	263.0	62	0.01	71.279624	65.0	1154	19.1	83	6.0	...	0	0	0	0	0	0	0	0	1	0
1	59.9	271.0	64	0.01	73.523582	62.0	492	18.6	86	58.0	...	0	0	0	0	0	0	0	1	0	0
2	59.9	268.0	66	0.01	73.219243	64.0	430	18.1	89	62.0	...	0	0	0	0	0	0	1	0	0	0
3	59.5	272.0	69	0.01	78.184215	67.0	2787	17.6	93	67.0	...	0	0	0	0	0	1	0	0	0	0
4	59.2	275.0	71	0.01	7.097109	68.0	3013	17.2	97	68.0	...	0	0	0	0	1	0	0	0	0	0

Tabel 2. Tabel dataset setelah dilakukan pembuangan variabel *dummy*

Setelah dilakukan pembuangan variabel *dummy* pada dataset, selanjutnya dapat dilakukan *splitting* menjadi data latih dan data uji.

```
#Untuk Data Regresi (life)
X_trainr,X_testr,Y_trainr,Y_testr=train_test_split(X_life,Y_life,test_size=0.3)
```

Gambar 4. *Splitting* menjadi data latih dan data uji

```
print(X_trainr.shape)
print(X_testr.shape)
print(Y_trainr.shape)
print(Y_testr.shape)
```

```
(2056, 226)
(882, 226)
(2056,)
(882,)
```

Gambar 5. Hasil *splitting*

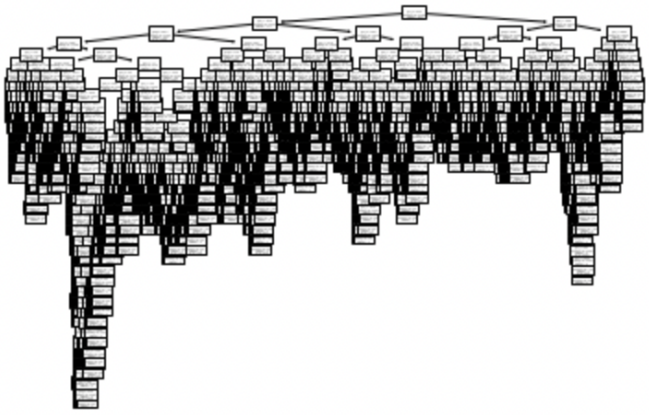
Dalam *splitting* menjadi data latih dan data uji dengan proporsi data uji sebesar 30% dari total observasi dan sisanya proporsi data latih. Berdasarkan pengujian didapatkan sebanyak 2056 observasi dari 226 variabel independen untuk data latih, dan didapatkan sebanyak 882 observasi dari 226 variabel independen untuk data uji. Selain itu juga didapatkan sebanyak 2056 observasi dari variabel dependen untuk data latih dan sebanyak 882 observasi dari variabel dependen untuk data uji. Selanjutnya dapat dilakukan pemodelan menggunakan dataset **Tabel 2** untuk analisis lebih lanjut.

4.3 *Modeling* atau Pemodelan

Pada pemodelan analisis regresi pada dataset akan dengan menggunakan algoritma *decision tree* dan *random forest*.

4.3.1 *Decision Tree*

Dibentuk algoritma *decision tree* atau pohon keputusan sehingga diperoleh plot *decision tree* sebagai berikut,



Gambar 6. Plot *decision tree*

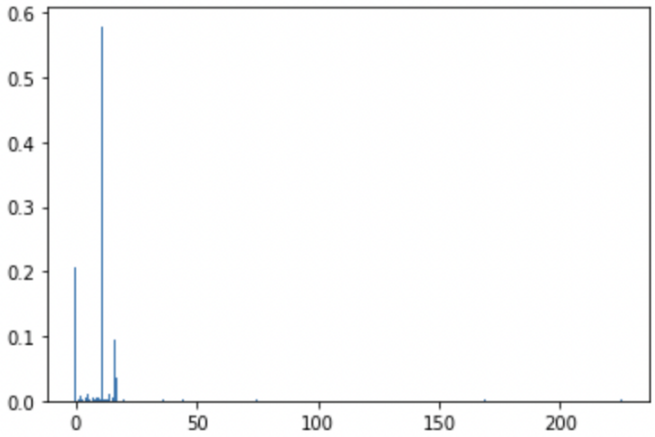
Berdasarkan **Gambar 6** terlihat bahwa plot dari algoritma *decision tree* atau pohon keputusan yang didapatkan dari hasil analisis pada dataset sangat rimbun. Dari algoritma *decision tree* tersebut didapatkan nilai performa model regresi *decision tree* pada dataset sebagai berikut,

```
Y_pred=clf1.predict(X_testr)
print("MSE:",mean_squared_error(Y_testr,Y_pred))
print("MAE:",mean_absolute_error(Y_testr,Y_pred))
```

```
MSE: 8.170712628424068
MAE: 1.6886624412971012
```

Gambar 7. Performa model regresi *decision tree* pada dataset

Berdasarkan **Gambar 7** diperoleh nilai MSE sebesar 8.170712628422068 dan nilai MAE sebesar 1.6886624412971012 sehingga dapat dikatakan bahwa performa model regresi menggunakan algoritma *decision tree* pada dataset cukup baik. Selanjutnya dilakukan pencarian variabel *importance* dari masing-masing variabel terhadap model.

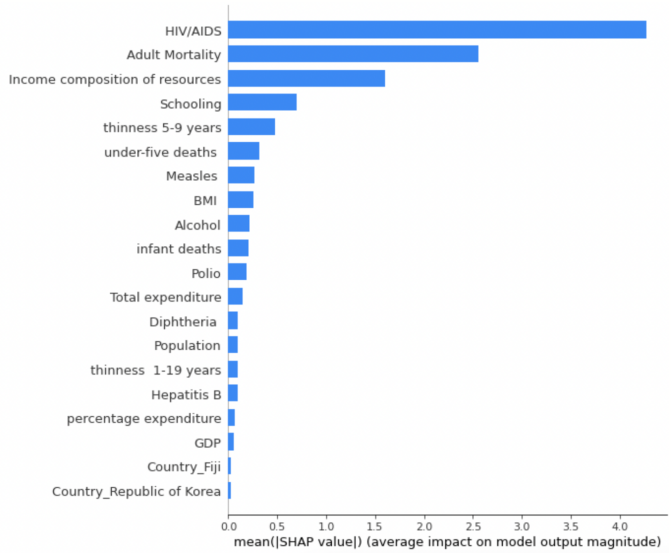


Gambar 8. Grafik variabel *importance* pada model regresi *decision tree*

```
Feature: 0, Score: 0.18791
Feature: 1, Score: 0.00329
Feature: 2, Score: 0.00650
Feature: 3, Score: 0.00159
Feature: 4, Score: 0.00288
Feature: 5, Score: 0.01079
Feature: 6, Score: 0.00659
Feature: 7, Score: 0.00920
Feature: 8, Score: 0.00511
Feature: 9, Score: 0.00539
Feature: 10, Score: 0.00171
Feature: 11, Score: 0.60665
Feature: 12, Score: 0.00150
Feature: 13, Score: 0.00337
Feature: 14, Score: 0.00349
Feature: 15, Score: 0.01122
Feature: 16, Score: 0.09482
Feature: 17, Score: 0.02829
Feature: 18, Score: 0.00004
Feature: 19, Score: 0.00000
Feature: 20, Score: 0.00032
```

Gambar 9. Nilai variabel *importance* pada model regresi *decision tree*

Berdasarkan **Gambar 8** dan **Gambar 9** terlihat bahwa variabel pada *Feature* atau kolom dengan nomor 0 dan 11 merupakan variabel yang memiliki nilai variabel *importance* tertinggi. Artinya kedua variabel tersebut merupakan variabel yang memiliki pengaruh paling signifikan terhadap model. Selain dari kedua gambar tersebut untuk melihat variabel yang memiliki pengaruh signifikan terhadap model dapat digunakan grafik berikut,



Gambar 10. Grafik *average impact on model*

Berdasarkan **Gambar 10** terlihat bahwa garis biru yang terbentuk pada grafik semakin ke bawah semakin pendek artinya semakin pendek garis yang dihasilkan, maka semakin kecil pengaruh variabel tersebut dalam memberikan pengaruh pada model. Sehingga dapat dikatakan bahwa variabel *HIV/AIDS* merupakan variabel yang memberikan pengaruh paling signifikan atau besar pada model. Kemudian disusul dengan variabel *Adult Mortality, Income composition of resources*, dan beberapa variabel lainnya.

4.3.2 *Random Forest*

Dibentuk algoritma *random forest* sehingga diperoleh nilai MSE dan MAE pada model sebagai berikut,

```
Y_pred=rf2.predict(X_testr)
print("MSE:",mean_squared_error(Y_testr,Y_pred))
print("MAE:",mean_absolute_error(Y_testr,Y_pred))
```

```
MSE: 3.844855102386966
MAE: 1.2279248935913176
```

Gambar 11. Performa model regresi random forest pada dataset

Berdasarkan **Gambar 11** diperoleh nilai MSE sebesar 3.844855102386966 dan nilai MAE sebesar 1.22792489935913176 sehingga dapat dikatakan bahwa performa model regresi menggunakan algoritma *random forest* pada dataset cukup baik. Selanjutnya dilakukan pencarian performa model terbaik dengan mencari susunan *hyperparameter model* dengan diberikan ruang pencarian *hyperparameter* tertentu.

```
params={'n_estimators':(100, 200, 400),'max_features':('auto',226//3)}
rf2a=RandomForestRegressor()
gs=GridSearchCV(rf2a,params)
gs.fit(X_trainr,Y_trainr)
```

```
GridSearchCV(estimator=RandomForestRegressor(),
             param_grid={'max_features': ('auto', 75),
                         'n_estimators': (100, 200, 400)})
```

Gambar 12. Susunan *hyperparameter* dengan metode *grid search*

Diberikan nilai ruang pencarian *hyperparameter*, yaitu:

- *max_features* : '*auto*' , 75
- *n_estimators* : 100, 200, 400

Selanjutnya dilakukan pemanggilan untuk nilai *max_features* dan *n_estimators* mana yang memberikan performa model terbaik dengan nilai ruang pencarian *hyperparameter* yang diberikan.

```
gs.best_estimator_
```

```
RandomForestRegressor(max_features=75, n_estimators=400)
```

Gambar 13. Nilai *max_features* dan *n_estimators* yang terpilih

Berdasarkan **Gambar 13** diperoleh nilai pencarian performa model terbaik dengan nilai *max_features* sebesar 75 dan nilai *n_estimators* sebesar 400. Selanjutnya dilakukan pengujian apakah performa model regresi menggunakan susunan *hyperparameter* dengan metode *grid search* tersebut benar-benar baik untuk digunakan atau tidak.

```
Y_pred=rf2b.predict(X_testr)
print("MSE:",mean_squared_error(Y_testr,Y_pred))
print("MAE:",mean_absolute_error(Y_testr,Y_pred))
```

```
MSE: 3.7127836859723122
MAE: 1.2114488196690296
```

Gambar 14. Performa model regresi dengan metode grid search

Berdasarkan **Gambar 14** diperoleh nilai MSE sebesar 3.7127836859723122 dan nilai MAE sebesar 1.2114488196690296 sehingga dapat dikatakan bahwa performa model regresi dengan susunan *hyperparameter* dengan metode *grid search* pada dataset cukup baik. Dibandingkan dengan model awal didapatkan nilai MSE dan MAE dari menggunakan susunan *hyperparameter* dengan metode *grid search* ini lebih kecil artinya model yang didapatkan dengan susunan *hyperparameter* dengan metode *grid search* lebih baik digunakan daripada model awal.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis regresi yang telah dilakukan didapatkan model pada *decision tree* nilai MSE sebesar 8.170712628422068 dan nilai MAE sebesar 1.6886624412971012, sedangkan model pada *random forest* (susunan *hyperparameter* dengan metode *grid search*) didapatkan nilai MSE sebesar 3.7127836859723122 dan nilai MAE sebesar 1.2114488196690296. Oleh karena nilai MSE dan MAE yang didapatkan pada keduanya, maka dapat diambil kesimpulan bahwa model pada *random forest* (susunan *hyperparameter* dengan metode *grid search*) merupakan model yang paling baik digunakan untuk dataset ini karena memiliki nilai MSE dan MAE lebih kecil.

5.2 Saran

Saran yang dapat kami berikan, yaitu dalam menentukan prediksi angka harapan hidup dari 193 negara di dunia berdasarkan data dari 22 variabel meliputi faktor kesehatan, ekonomi, dan sosial dengan menggunakan analisis regresi linear berganda dapat dilakukan dengan menggunakan model pada *random forest* (susunan *hyperparameter* dengan metode *grid search*). Karena berdasarkan hasil prediksi tersebut didapatkan model yang memiliki tingkat akurasi tinggi.

DAFTAR PUSTAKA

Anggraini, Eviana, dan U. Listyaningsih. (2013). *Disparitas Spasial Angka Harapan Hidup Di Indonesia Tahun 2010*. Jurnal Bumi Indonesia, vol. 2, no. 3.

The World Bank. (n.d.). *Life expectancy at birth, total (years)*. worldbank.org. Diakses 20 Mei 2022 pada https://data.worldbank.org/indicator/SP.DYN.LE00.IN?end=2017&start=2017&view=bar

Maryani, H. dan L. Kristiana. (2017). *Pemodelan Angka Harapan Hidup (AHH) Laki-Laki dan Perempuan di Indonesia Tahun 2016*. Dinkes Bulettin, vol. 21, no. 2.

Suliztia, Mega Luna. (2020), *Penerapan Analisis Random Forest pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask*. Universitas Islam Indonesia.

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed. Canada: O'Reilly Media.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer

Breiman, L. (2001). *Random Forests. Machine Learning*, vol. 45, pp. 5-32.