LAPORAN PRAKTIKUM MACHINE LEARNING
"KLASIFIKASI UMUR ABALONE BERDASARKAN UKURAN FISIK DENGAN RANDOM FOREST"

Disusun Oleh :

Nama : Rosa Julia Erizka

NIM : 09011282126105

Kelas : SK6C

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

TAHUN 2024

DAFTAR ISI

BAB I
PENDAHULUAN

1.1. Latar Belakang

Salah satu spesies laut yang sangat menguntungkan, abalone digunakan dalam industri perikanan dan kuliner di seluruh dunia. Overfishing seringkali mengancam populasi abalone karena tingginya permintaan dan nilai komersilnya. Oleh karena itu, sangat penting untuk memprioritaskan pengelolaan sumber daya abalone yang berkelanjutan, yang mencakup peraturan yang mengatur ukuran minimal penangkapan dan periode penangkapan yang diperbolehkan. Kemampuan untuk menentukan umur abalone secara akurat dan efektif adalah komponen penting dalam manajemen ini.

Secara tradisional, umur abalone ditentukan melalui analisis cincin pertumbuhan pada cangkangnya; metode ini memerlukan banyak waktu dan sumber daya, dan seringkali subjektif dan tidak akurat. Akibatnya, ada kebutuhan akan metode yang lebih canggih dan objektif untuk menentukan umur abalone. Dalam hal ini, teknologi pembelajaran mesin memiliki banyak peluang yang luar biasa. Khususnya, algoritma Random Forest, yang telah terbukti berhasil menangani klasifikasi data yang sangat besar dengan tingkat akurasi yang tinggi, merupakan salah satu contohnya.

Tujuan penelitian ini adalah untuk membuat model klasifikasi umur abalone yang didasarkan pada ukuran fisiknya. Diharapkan bahwa dengan menggunakan model ini, akan ditemukan cara yang lebih objektif dan efisien untuk mengestimasi umur abalone, yang pada gilirannya akan mendukung kebijakan pengelolaan yang lebih tepat dan berkelanjutan. Penelitian ini juga diharapkan akan memberikan wawasan baru tentang penggunaan teknologi pembelajaran mesin dalam ilmu kelautan, khususnya dalam penelitian dan pengelolaan sumber daya kelautan.

1.2. Rumusan Masalah

Dalam penelitian ini, menggunakan metode Random Forest untuk mengklasifikasikan umur abalone berdasarkan ukuran fisiknya. Penelitian ini bertujuan untuk mempelajari hubungan antara ukuran fisik abalone dan umurnya. Metode tradisional yang umum digunakan untuk menentukan umur abalone, yaitu analisis cincin pertumbuhan pada cangkangnya. Metode ini sering kali memerlukan interpretasi manual yang subjektif dan rentan terhadap kesalahan, yang menimbulkan

pertanyaan tentang keakuratannya. Memverifikasi seberapa efektif data ukuran fisik ini dalam memprediksi umur adalah penting. Ini juga penting untuk mengetahui apakah ada pola atau hubungan non-linear yang belum ditemukan oleh metode konvensional. Akhirnya, kami ingin mengetahui apakah metode Random Forest, dibandingkan dengan metode konvensional, dapat lebih akurat dan efisien dalam mengklasifikasikan umur abalone.

1.3.Tujuan

Tujuan utama dari laporan ini adalah untuk mengembangkan sebuah model klasifikasi yang dapat menggunakan algoritma Random Forest untuk memprediksi usia abalone berdasarkan ukuran fisiknya. Model ini akan menggunakan variabel seperti panjang, tinggi, berat total, berat kering, berat isi, dan berat cangkang untuk menentukan kategori usia abalone. Kinerja model akan diukur dengan metrik seperti akurasi, presisi, recall, dan skor F1, yang akan membantu menentukan usia abalone yang sebenarnya. Tujuan dari penelitian ini adalah untuk menentukan metode terbaik untuk mengklasifikasikan usia abalone yang dapat digunakan dalam industri perikanan dan penelitian biologi laut.

BAB II
TINJAUAN PUSTAKA

2.1. Algoritma Random Forest

Random forest classifier adalah metode klasifikasi yang terdiri dari kumpulan pohon keputusan yang nantinya akan dijadikan suara untuk mendapatkan hasil terakhir dari pendeteksian sarkasme dengan pendukung berupa data latihan dan fitur acak yang independen dengan berbagai karakteristik. Pohon keputusan dibuat dengan menentukan node akar dan berakhir dengan node beberapa daun. Random Forest berasal dari metode CART (Classification and Regression Trees), yang juga merupakan algoritma dari teknik pohon keputusan. Yang membedakan metode Random Forest dari metode CART adalah bahwa metode Random Forest menggunakan metode aggregating bootstrap (bagging) dan juga menggunakan seleksi fitur random, yang juga disebut sebagai seleksi fitur random. Random Forest adalah kombinasi dari metode pohon keputusan yang ada yang digabungkan dan digabungkan ke dalam suatu model. Metode Random Forest memiliki tiga prinsip utama:

(1) Membangun pohon prediksi menggunakan sampling bootstrap;
(2) Memprediksi tiap pohon keputusan dengan prediktor acak; dan
(3) Random Forest membuat prediksi dengan menggabungkan hasil dari tiap pohon keputusan dengan suara mayoritas untuk klasifikasi atau rata-rata untuk regresi.

2.2. Keunggulan Algoritma Random Forest

Keunggulan algoritma Random Forest dalam klasifikasi antara lain:

1. Tahan terhadap overfitting karena menggunakan banyak teknik averaging dan pohon keputusan.
2. Dapat mengatasi dataset yang besar dengan banyak fitur karena mampu menangani banyak variable secara efisien
3. Random Foerst mampu mengatasi missing value bahkan tanpa preprocessing data yang rumit.
4. Dapat digunakan pada data numerik maupun kategorikal.
5. Dapat melakukan pemilihan fitur secara otomatis.

2.3. Kekurangan Algoritma Random Forest

Kekurangan algoritma Random Forest dalam klasifikasi anatara lain:

1. Membutuhkan waktu dan sumberdaya yang lebih besar dan kompleks daripada metode yang lainnya.
2. Jika jumlah pohon pada model terlalu besar maka akan menyebabkan overfitting pada data training.
3. Tidak cocok untuk dataset yang bersifat dinamis atau berubah atau mengalami perubahan seiring waktu (drift)
4. Memerlukan optimasi pada parameternya
5. Tidak cocok untuk data yang memiliki kelas imbalance/tidak seimbang.

BAB III
PEMBAHASAN

3.1. Pemilihan Topik dan Dataset

3.1.1. Pemilihan Topik

Topik yang dipilih adalah Pengklasifikasian Usia Abalone Berdasarkan Physical Measurements Dengan Metode Random Forest. Projek ini akan dilakukan pengelompokkan usia pada kerang abalone yang dilihat dari pengukuran fisik abalone tersebut (termasuk ring, panjang shell, diameter, berat dsb). Dimana, pengelompokan usia ini dikelompokkan menjadi tiga yaitu 'Young', 'Middle Age', dan 'Old'. Pengklasifikasian usia abalone ini bertujuan melihat persebaran populasi dan menekan eksploitas berlebih sehingga menjaga eksistem laut dengan baik.

3.1.2. Plan Dataset dan Pemilihan Dataset

Disini menggunakan dataset dari UCI berikut link-nya : https://archive.ics.uci.edu/dataset/1/abalone dari dataset tersebut terdapat 10 Variabel hasil pengukuran fisik abalone. Dari pengukuran ini lah akan dikelompokkan usia abalone tersebut dimana variabel dependen berupa variabel Rings dimana variabel ini menjelaskan bahwa semakin besar Rings pada abalone maka semakin tua abalone tersebut, karena Rings pada abalone akan bertambah semakin bertambahnya usia abalone tersebut.

3.2. Explorasi dan Pemahaman Data

3.2.1. Analisis Eksplorasi Dataset

Berikut dataset yang digunakan :

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7
...	...	...	...	...	...	...	...	...	...
4172	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11
4173	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10
4174	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9
4175	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10
4176	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12

4177 rows × 9 columns

Dataset yang digunakan diambil pada UCI Machine Learning Repository. Dari pemaparan dataset diatas, terdapat 9 variabel dari pengukuran fisik abalone dengan isi data sebanyak 4177 baris. Dari ke-sembilan variabel diatas yaitu : Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight, dan Rings memiliki peranan penting dalam menentukan umur abalone. Dimana variabel Rings merupakan variabel dependen, dan fitur Sex, Length, Diameter, and Height, Whole weight, Shucked weight, Viscera weight, Shell weight merupakan variabel independennya.

```
RangeIndex: 4177 entries, 0 to 4176
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Sex             4177 non-null   object
 1   Length          4177 non-null   float64
 2   Diameter        4177 non-null   float64
 3   Height          4177 non-null   float64
 4   Whole weight    4177 non-null   float64
 5   Shucked weight  4177 non-null   float64
 6   Viscera weight  4177 non-null   float64
 7   Shell weight    4177 non-null   float64
 8   Rings           4177 non-null   int64
dtypes: float64(7), int64(1), object(1)
memory usage: 293.8+ KB
```

Dataset yang digunakan memiliki 9 variabel dan tidak terdapat nilai null ditiap variabelnya. Tipe data pada tiap variabel :

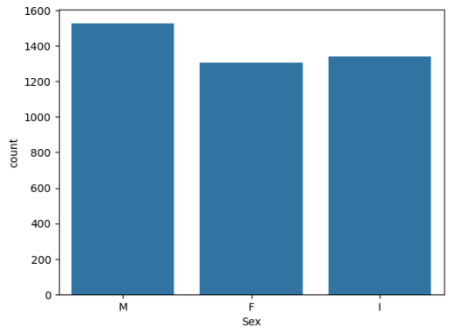
1. Tipe data object, terdapat pada variabel Sex dengan 3 nilai (M = Male, F = Female, dan I= Infant).
2. Tipe data float/continuous, terdapat pada variabel Length, Diameter, and Height, Whole weight, Shucked weight, Viscera weight, Shell weight. Dengan skala pengukurun milimeter (mm) dan gram (g) pada pengukuran berat.
3. Tipe data Integer, terdapat pada variabel Rings yang memuat berapa banyak rings pada abalone dimana jumlah ring +1.5 merupakan umur dari abalone tersebut.

Penjelasan Tiap Variabel

1. Sex

   Variabel ini menunjukkan jenis kelamin abalone, terdapat tiga sex pada abalone yaitu: M = Male, F = Female, dan I = Infant/bayi. Persebaran jumlah pada dataset diatas pada variabel Sex ada pada diagram dibawah, dimana Male berjumlah lebih banyak daripada Female dan Infant. dimana Male berjumlah 1528, Female : 1342, Infant : 1307.

```
M    1528
I    1342
F    1307
Name: Sex, dtype: int64
```



2. Length (Panjang Cangkang)

   Variabel ini berisikan pengukuran panjang cangkang kerang, dengan satuan ukur milimeter (mm). Dari describe diatas bahwa ukuran terpendek berupa 0.075 mm dan ukuran terpanjang 0.815 mm

3. Diameter

   Variabel ini berisikan pengukuran dari diameter pada abalone dengan pengukuran tegak lurus dengan panjang/lenght abalone. Satuan ukur berupa (mm) dengan ukuran diameter paling kecil 0.055 mm dan paling besar 0.65 mm.

4. Height (Tinggi)

   Variabel ini berisikan pengukuran dari Tinggi cangkang termasuk daging di dalam pada abalone. Satuan ukur berupa (mm) dengan ukuran abalone paling tinggi sebesar 1.13 mm.

5. Whole Wight (Berat Keseluruhan Abalone)

   Variabel ini berisikan pengukuran dari berat keseluruhan pada abalone termasuk berat cangkang dan daging dalam abalone. Satuan ukur berupa (gram) dengan ukuran paling berat 2.8255 g dan paling ringan 0.002 g.

6. Shucked Weight (Berat Daging)

   Variabel ini berisikan pengukuran dari berat daging pada abalone yang telah dikeluarkan dari cangkangnya. Satuan ukur berupa (gram) dengan yang paling berat sebesar 1.488 g dan paling ringan 0.001 g.

7. Viscera Weight (Berat Usus)

   Variabel ini berisikan pengukuran dari berat usus pada abalone yang telah dipisahkan dari cangkang dan dikeluarkan darah pada abalone. Satuan ukur berupa (gram) dengan ukuran berat usus paling kecil 0.0005 g dan paling besar 0.76 g.

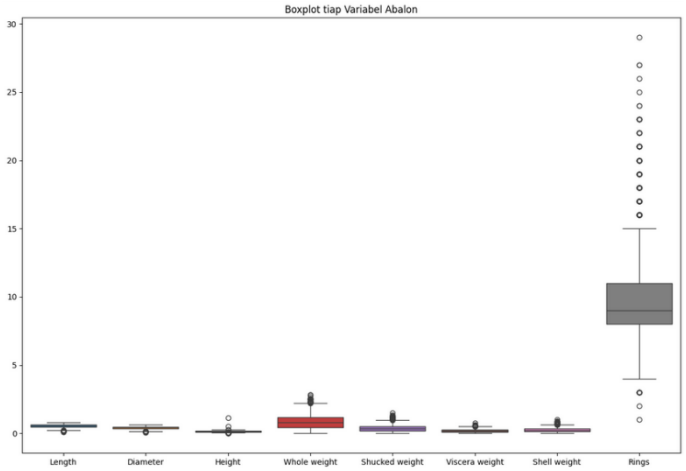
8. Shell Weight (Berat Cabgkang)

   Variabel ini berisikan pengukuran dari berat pada cangkang abalone yang telah dikeringkan terlebih dahulu. Satuan ukur berupa (gram) dengan ukuran diameter paling kecil 0.055 mm dan paling besar 0.65 mm.

9. Rings

   Variabel ini berisikan pengukuran dari jumlah cincin pada abalone. Dimana penentuan umur pada abalone itu dilakukan dengan menjumlahkan jumlah rings pada balone dengan +1.5. Dengan jumlah rings paling sedikit adalah 1, maka perkiraan usian abalon tersebut : 1 + 1.5 = 2.5 tahun, dan jumlah rings paling banyak yaitu 29, maka usian abolen tersebut 29 + 1.5 = 30.5 tahun.

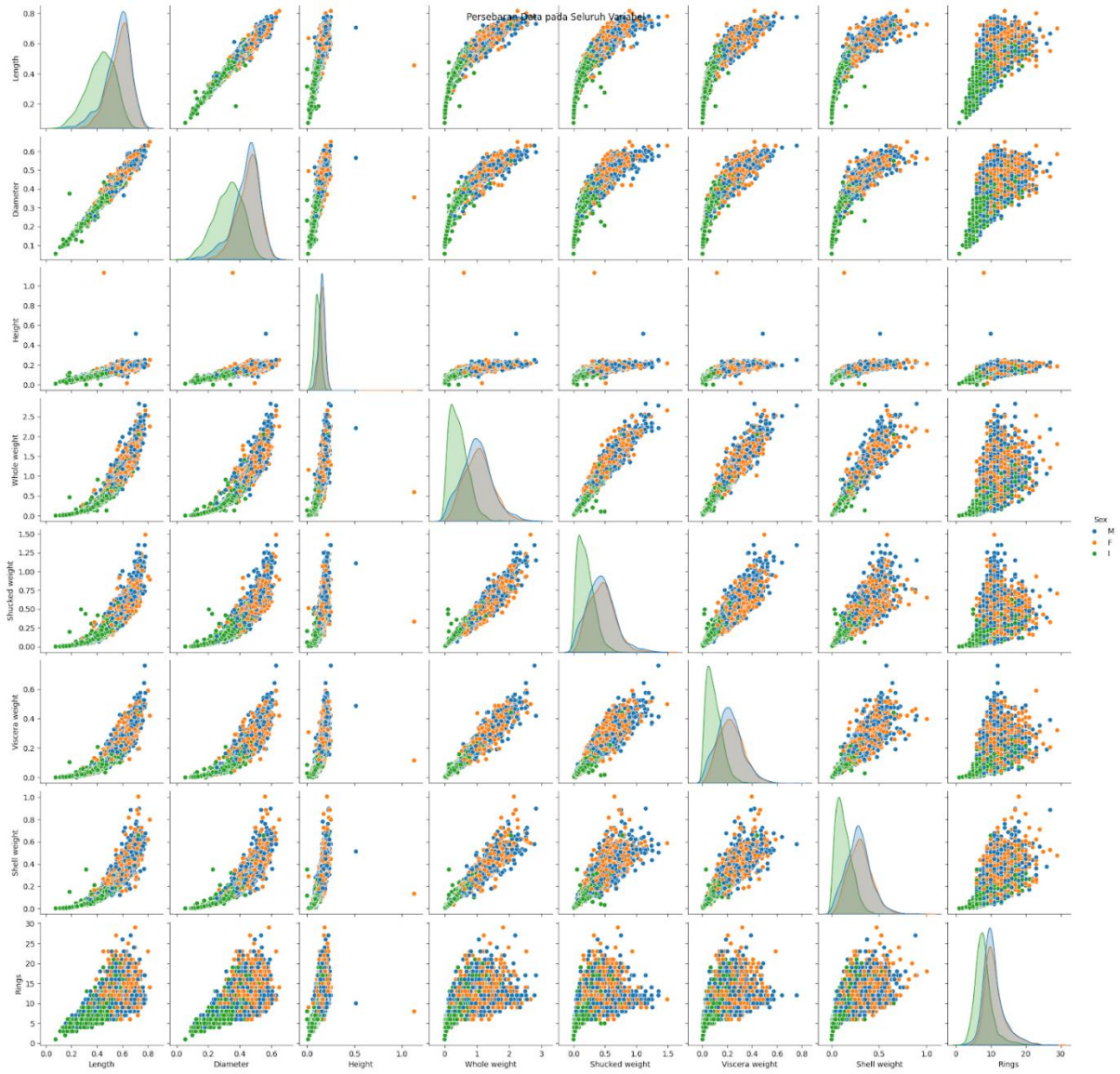
Boxplot Pada Tiap Variabel:



Dapat dilihat pada boxplot diatas :

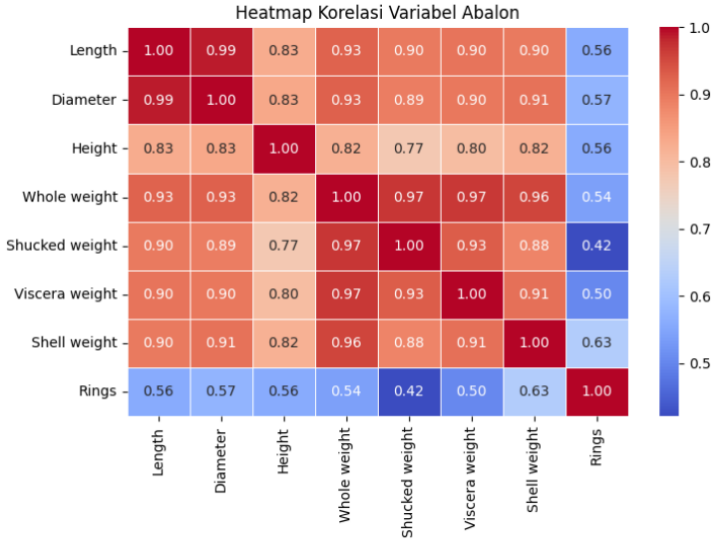
a. Pada variabel Lenght terdapat beberapa outlier yang mendekati nilai 0, dimana rata-rata isi dari variabel lenght hanya 0.52 saja jadi pesebaran data diantara nilai terkecilnya 0.075 s.d nilai terbesarnya 0.815
b. Pada variabel Diameter juga terdapat beberapa outlier dengan nilai mendekati 0, karena pesebaran nilai variabelnya diantara nilai terkecilnya 0.055 s.d nilai terbesarnya 0.65
c. Pada Variabel Height terdapat outlier diatas nilai 1.4 dan mendekati 0, dimana pesebaran data pada variabel ini jika dilihat dari median nya berada pada nilai 1.4
d. Pada variabel Whole Weight, Shucked Weight, Viscera Weight dan Shell Weight juga memiliki nilai outlier diatas kotak/whicker
e. Pada Variabel Rings tjuga terdapat outlier yang cukup berjauhan diatas 15 dan dibawah 4

Persebaran Data pada tiap Variabel :



Dara pairplot dilihat persebaran pada masing-masing variabel. Dengan warna biru untuk Male, warna orange untuk Female, dan warna hijau untuk Infant. Pada variabel Lenght/panjang, semakin panjang ukuran cangkang maka akan berbanding lurus dengan ke 8 variabel lainnya. Begitu pula pada tiap variabel lainnya saling berbanding lurus, yang artinya jika pajang abalone semakin besar maka variabel seperti diameter, berat akan semakin besar pula. Dapat dilihat juga, bahwa abalone dengan Sex Infant/Bayi memiliki ukuran fisiknya domoninan lebih kecil ditiap variabel pengukuran fisik abalone.

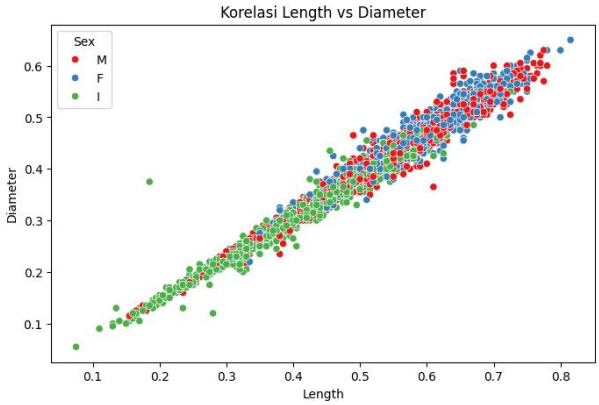
Korelasi ditiap variabel dengan heatmap :



Pada korelasi di heatmap, jika mendekati -1 berarti korelasi negatif, mendekati +1 korelasi postif, dan jika 0 maka tidak ada korelasi. Dari heatmap diatas variabel yg memiliki niali korelasi yang paling tinggi itu adalah variabel length/panjang dengan variabel diameter dengan nilai korelasi 0.99. Yang artinya kedua variabel ini memiliki korelasi positif, dimana jika panjang cangkang kerang semakin panjang maka diameternya pun akan semakin besar pula.

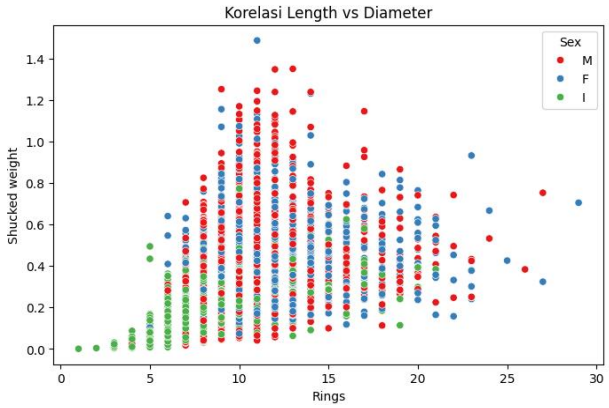
Hubungan Korelasi pada Variabel

a. Korelasi Variabel Length dan Diameter



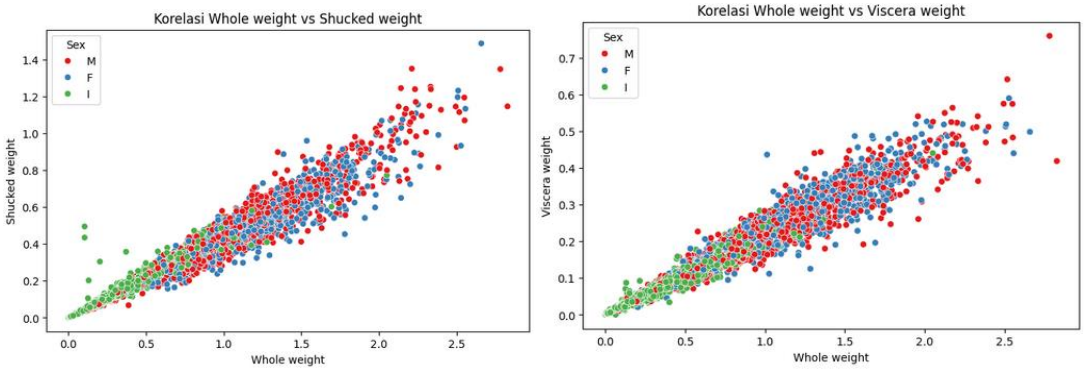
Memiliki korelasi yang tinggi artinya saling berkorelasi. Jika semaking panjang cangkang abalone maka akan semakin besar pula diameter pada abalone. Dapat dilihat bahwa, semakin besar nilai Lenght makan semakin besar pula nilai variabel Diameter. Dari scatterplot ini, dapat dilihat bahwa Infant/Bayi mendominasi nilai yang kecil dari kedua variabel ini. Karena Infant masih dalam tahap pertumbuhan.

b. Korelasi Variabel Ring dan Shucked Weight



Memiliki korelasi yang rendah diantara variabel lainnya. Pada scatterplot ini, sedikit sulit untuk melihat korelasi antara variabel Shucked weight dan Rings ini. Tapi masih dapat dilihat bahwa Infant/bayi masih mendominasi nilai terkecil karena masih dalam tahap pertumbuhan.

c. Korelasi Variabel Whole Weight dengan Viscera Weight dan Shucked Weight



Memiliki nilai korelasi yang sama sebesar 0.97. Dapat dilihat pada kedua scatterplot yang menampilkan variabel Whole Weight dengan Viscera Weight dan Shucked Wight, karena memiliki korelasi yang tinggi sebesar 0.97 ketiga variabel ini saling berkorelasi positif. Karena jika berat keseluruhan abalone Whole weight besar, maka kedua variabel Shucked Weight/ berat daging dan Viscera Weight/berat usus juga akan menjadi berat juga. Berlaku kebalikannya juga, jadi jika berat daging dan usus nya besar maka berat keselurahan juga akan semakin berat.

3.2.2. Identifikasi Masalah Dataset

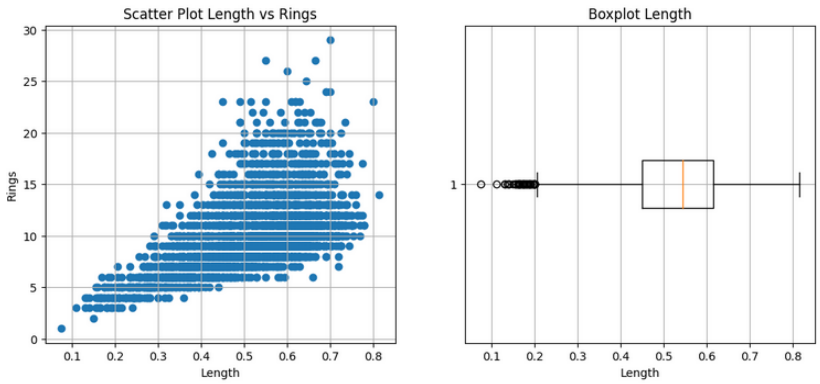
Pada dataset yang digunakan akan dilakukan pembersihan Outlier pada tiap variabelnya. Dengan menggunakan code df.isnull().sum() mengahasilkan bahwa dataset yang digunakan tidak memilikinilai null, sebelum dilakukan data training maka dilakukan proses pembersihan data terlebih dahulu dari nilai outliers pada tiap variabel yang ada pada dataset. Pada fitur Sex dilakukan encoding dengan get_gummies pada data categorical Sex, agar dapat diolah selanjutnya untuk pembuatan model. Dengan get_dummies dapat memisahkan label apa saja yang ada pada variabel Sex menjadi Sex_F (untuk Female), Sex_I (Infant) dan Sex_M (Male).

```
df = pd.get_dummies(df)
df.head()
```

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings	Sex_F	Sex_I	Sex_M
0	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15	0	0	1
1	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7	0	0	1
2	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9	1	0	0
3	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10	0	0	1
4	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7	0	1	0

Dan pada fitur Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight dilakukan preprocessing dengan menghapus niali outlier yang dilihat dari boxplot dan pairplot untuk melihat pesebaran datanya.

a. Fitur Length :



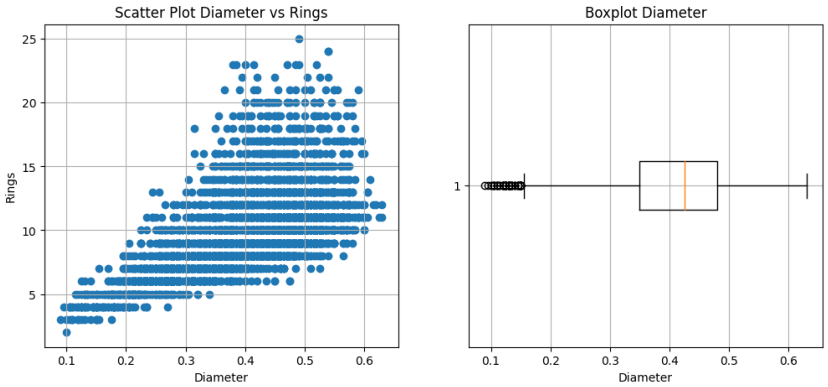
Dari scatter plot diatas terlihat pesebaran data Length terhadap Rings yang menjadi data target, dan pada boxplot terlihat ada outlier di bawah 0.2 yang akan dilakukan penghapusan seperti dibawah berikut ini :

```
df.drop(df[(df['Length'] <0.1) &(df['Rings'] < 5)].index, inplace = True)
df.drop(df[(df['Length']<0.8) & (df['Rings'] > 25)].index, inplace = True)
df.drop(df[(df['Length']>=0.8) & (df['Rings'] < 25)].index, inplace = True)
```

Pada fitur Length dilakukan penghapusan pada data :

- Nilai length yang kurang dari 0.1 dan Rings kurang dari 5
- Nilai length yang kurang dari 0.8 dan Rings lebih dari 25
- Nilai length yang lebih dari 0.8 dan Rings kurang dari 25

b. Fitur Diameter

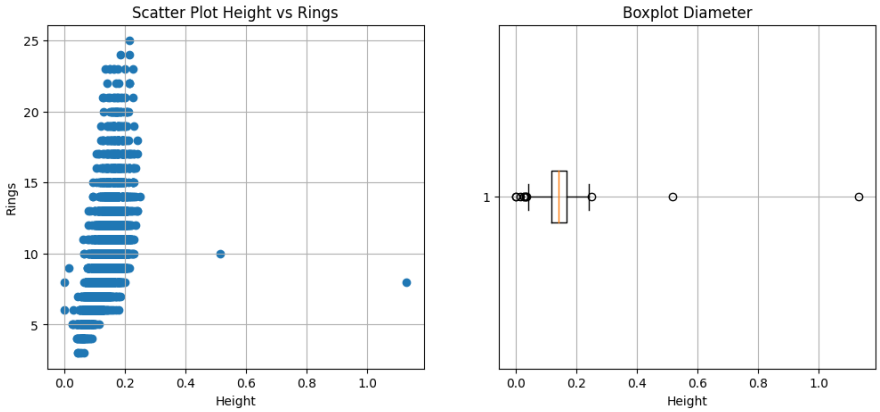


Pada fitur Diameter diatas dilakukan penghapusan pada data :

- Nilai Diameter yang kurang dari 0.15 dan Rings kurang dari 5
- Nilai Diameter yang kurang dari 0.6 dan Rings lebih dari 25
- Nilai Diameter yang lebih dari 0.6 dan Rings kurang dari 25

Sehingga jika dilihat hasil shape pada df memuliki baris sebanyak 4120 data.

c. Fitur Height

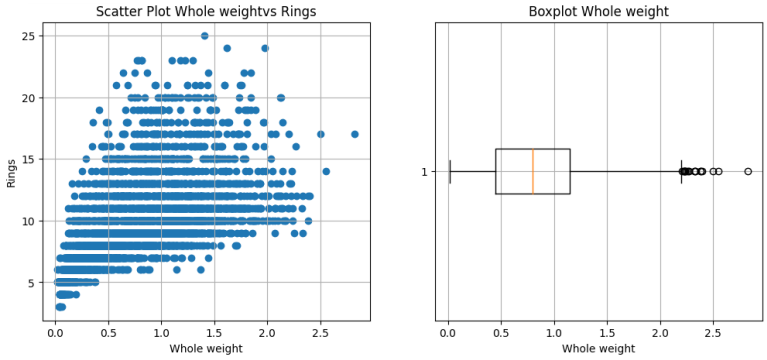


Pada fitur Height diatas dilakukan penghapusan pada data :

- nilai Height yang lebih dari 0.4 dan Rings kurang dari 15
- nilai Height yang kurang dari 0.4 dan Rings lebih dari 25

Sehingga jika dilihat hasil shape pada df memiliki baris sebanyak 4118 data.

d. Fitur Whole_Weight

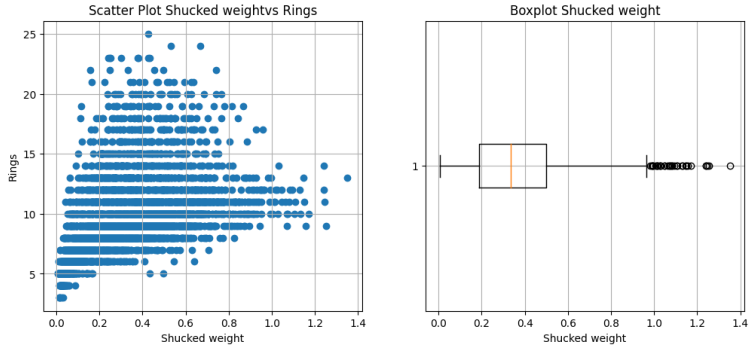


Pada fitur whole weight diatas dilakukan penghapusan pada data :

- nilai Whole weight yang kurang dari 2.5 dan Rings lebih dari 15
- nilai Whole weight yang lebih dari 2.5 dan Rings kurang dari 25

Sehingga jika dilihat hasil shape pada df memiliki baris sebanyak 4116 data

e. Fitur Shucked_Weight

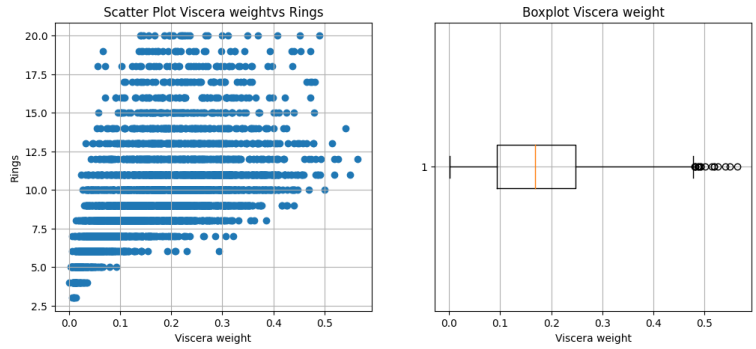


Pada fitur Shucked weight diatas dilakukan penghapusan pada data :

- nilai Shucked weight yang lebih dari 1 dan Rings kurang dari 20
- nilai Shucked weight yang kurang dari 1 dan Rings lebih dari 20

Sehingga jika dilihat hasil shape pada df memiliki baris sebanyak 4116 data

f. Fitur Viscera_Weight

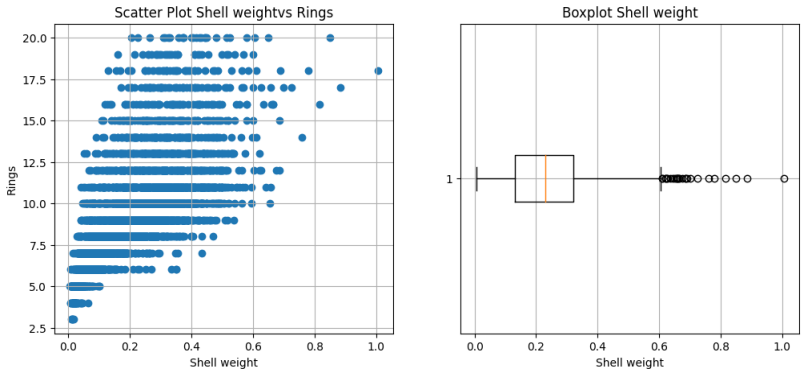


Pada fitur Viscera weight diatas dilakukan penghapusan pada data :

- nilai Viscera weight yang lebih dari 0.5 dan Rings kurang dari 20
- nilai Viscera weight yang kurang dari 0.5 dan Rings lebih dari 25

Sehingga jika dilihat hasil shape pada df memiliki baris sebanyak 4048 data

g. Fitur Shell_weight



Pada fitur Shell Weight diatas dilakukan penghapusan pada data :

- nilai Shell weight yang lebih dari 0.6 dan Rings kurang dari 25
- nilai Shell weight yang kurang dari 0.8 dan Rings lebih dari 25

Sehingga jika dilihat hasil shape pada df memiliki baris sebanyak 4018, yang awalnya 4177. Hal ini dilakukan pada semua fitur yang disebutkan tadi. menghasilkan data yang telah bersih dengan berisikan 2018 baris dengan 11 kolom yang dimasukkan pada variabel data :

```
data = df.copy()
data
```

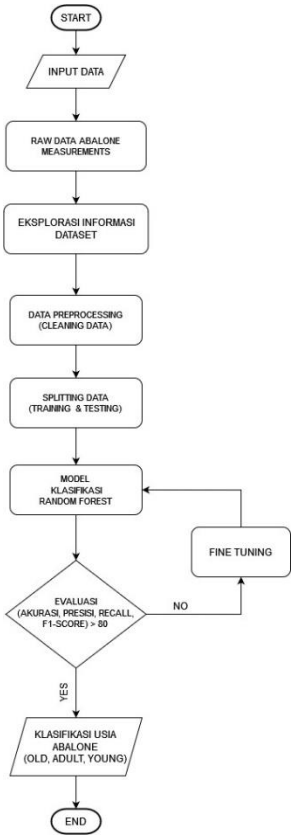
	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings	Sex_F	Sex_I	Sex_M
0	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15	0	0	1
1	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7	0	0	1
2	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9	1	0	0
3	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10	0	0	1
4	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
4172	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11	1	0	0
4173	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10	0	0	1
4174	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9	0	0	1
4175	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10	1	0	0
4176	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12	0	0	1

4018 rows × 11 columns

3.3. Pemilihan Model

3.3.1. Alur Kerja Algoritma

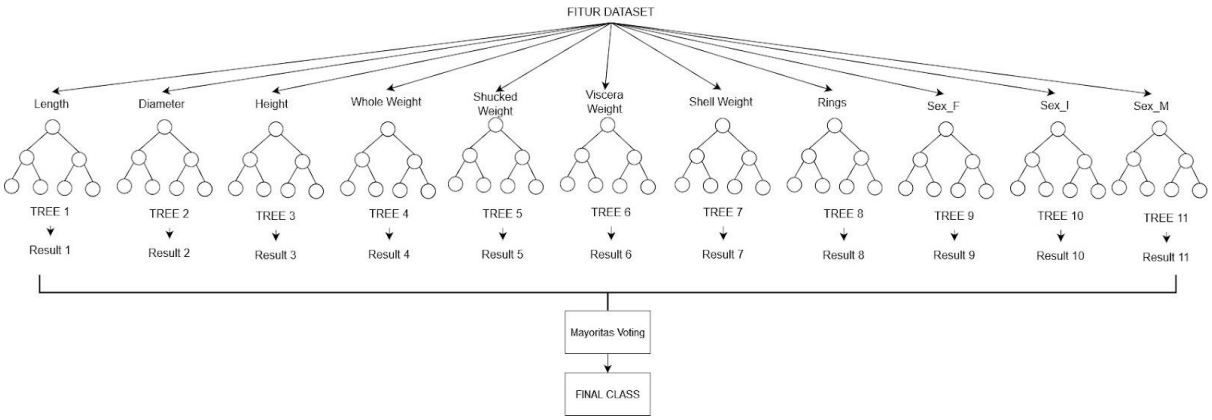
Menggunakan Supervised Learning dengan metode klasifikasi Random Forest untuk melakukan pengkalasifikasian pada umur abalone, menggunkanan dataset abalone yang telah dilakukan preprocessing terlebih dahulu.Berikut alur kerja pada projek yang dibuat :



Pertama, data dilakukan analisis/ekplorasi terlebih dahulu untuk melihat korelasi dan informasi pada data. Yang kemudian akan dilakukan data preprocessing dari data raw menjadi data yang siap dilakukan training dan testing. Kemudian, data dibagi menjadi data training dan testing yang akan dimasukkan ke dalam model pengklasifikasian Random Forest. Hasil dari pengklasifikasian akan dievaluasi menggunakan performence matrics.

3.3.2. Arsitektur Algoritma Random Forest

Berikut arsitektur yang digunakan yaitu Random Forest untuk melakukan klasifikasi pada usia abalone seperti gambar dibawah :



Dapat dilihat data fitur yang digunakan ada 11 yaitu Sex (Sex_F, Sex_I,Sex_M), Length, Diameter, Height, Whole_weight, Shucked_weight, Viscera_weight, dan Shell_weight. Dari data fitur ini lah akan dilakukan klasifikasi terhadap umur abalone dengan variabel targetnya Rings.

3.4. Pelatihan dan Validasi Model

3.4.1. Splitting Data dan Standarization

Setelah dilakukan data preprocessing, maka dilakukan pemisahan data X dan data y yang menjadi target, yang kemudian akan dilakukan standarisasi data agar rentang data pada tiap variabel tidak ada yang mendominasi, dan dilakukan splitting data menjadi data train dan data testing :

```
#membagi dataset menjadi varibel X yang merupakan varabel fitur, dan y yang menjadi variabel target
X = data.drop('Rings', axis = 1)
y = pd.cut(data['Rings'], bins=[0, 6, 13, 30], labels=['Young', 'Adult', 'Old'])
```

Dari codingan di atas :

- variabel X berisi fitur : Sex (Sex_F, Sex_I,Sex_M), Length, Diameter, Height, Whole_weight, Shucked_weight, Viscera_weight, dan Shell_weight.
- variabel y berisi target : Yaitu Rings, dimana diabagi menjadi 3 bagian (0-6 untuk Young, 7-13 untuk Adult, dan 14-29 untuk Old)

Lalu melakukan standarisasi data agar rentang nilai pada variabel tidak ada yang mendominasi sehingga dapat menghasilkan performas model yang baik. Disini menggunakan StandarScaler dengan melakukan fit_transform pada variabel X.

```
standardScale = StandardScaler()
standardScale.fit_transform(X)
```

```
array([[-0.58943755, -0.43746459, -1.18790132, ..., -0.67239438,
        -0.69591664,  1.32650495],
       [-1.51363268, -1.50104896, -1.32537216, ..., -0.67239438,
        -0.69591664,  1.32650495],
       [ 0.07070184,  0.14750682, -0.08813461, ...,  1.48722242,
        -0.69591664, -0.75386074],
       ...,
       [ 0.68683193,  0.73247822,  1.83645712, ..., -0.67239438,
        -0.69591664,  1.32650495],
       [ 0.90687839,  0.83883666,  0.3242779 , ...,  1.48722242,
        -0.69591664, -0.75386074],
       [ 1.65503635,  1.58334572,  1.56151545, ..., -0.67239438,
        -0.69591664,  1.32650495]])
```

Dengan SelectKBest untuk membantu memilih fitur yang terbaik dari X dan y, yang kemudian akan ditranform ke dalam variabel KBest, dan dilakukan spliting data dengan perbandingan train dan testing sebesara 70% : 30%

```
selectkBest = SelectKBest()
KBest = selectkBest.fit_transform(X, y)

X_train, X_test, y_train, y_test = train_test_split(KBest, y, test_size = 0.3)
```

SelectKBest digunakan untuk memilih fitur yang paling relevan dan terbaik untuk dimasukkan kedalam model. Disini dilakukan pemisahan data train dan data test dengan data train sebesar 70% dan data set sebesar 30% dari hasil pemilihan fitur oleh SelctKBest dan dari y.

3.4.2. Training Model Random Forest

Pada pembangunan model klasifikasi Random Forest ini, menggunakan pohon keputusan atau n_estimator sebanyak 150, max_depth 10, dengan random_state 100 untuk hasil trainingnya tetap.

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import accuracy_score, classification_report

# Membuat model klasifikasi random forest:
rf1 = RandomForestClassifier(n_estimators=150, max_depth=10, random_state=100)

# training data pada model
rf1.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(max_depth=10, n_estimators=150, random_state=100)
```

Pembuatan model ditunjukkan pada code :

model = RandomForestClassifier(n_estimators=150, max_depth=10, random_state=100)

Dimana menggunakan klasifikasi Random Forest, dengan banyak pohon keputusan yang digunakan sebesar 150, dan nilai random_state agar nilai training tidak berubah sebesar 100.

Training model dilakukan pada code : model.fit(X_train, y_train), dimana model yang telah dibuat dilakukan fit pada X_train dan y_train.

Berikut hasil evaluasi dari model yang telah dibuat di tampilkan pada classification report :

```
Random Forest Training Accuracy:  0.914651493598862
Random Forest Test Accuracy:  0.8266998341625207

Classification Report:
              precision    recall  f1-score   support

       Adult       0.83      0.98      0.90       924
         Old       0.85      0.15      0.26       143
       Young       0.80      0.53      0.64       139

    accuracy                           0.83      1206
   macro avg       0.82      0.55      0.60      1206
weighted avg       0.83      0.83      0.79      1206
```

Dapat dilihat perhitungan akurasi dari y_test dan y_pred (didapatkan dari melakukan prediksi dengan model yang telah dibuat pada X_test) menghasilkan nilai akurasi training mengunakan klasifikasi dengan model random forest dengan memakai n_estimators sebanyak 150 dengan random_state 100, sebesar 91.465 dengan akurasi testing 82.669. Pada laporan klasifikasi diatas menjelaskan nilai presisi, recall, dan f1-score pada ketiga kelas diatas yaitu Old, Adult, dan Young. Dengan nilai matrik evaluasi pada tiap kelas sebesar:

1. Adult : memiliki presisi sebesar 0.83, recall 0.98, dan f1-score sebesar 0.90
2. Olds : memiliki presisi sebesar 0.85, recall 0.15, dan f1-score sebesar 0.26
3. Young : memiliki presisi sebesar 0.80, recall 0.53, dan f1-score sebesar 0.64

Dimana presisi untuk mengukur seberapa tepatnya model yang telah dibuat melakukan prediksi terhadap dataset asli, recall digunakan untuk mengukur seberapa baik model random forest diatas mampu menemukan semua instance positif, sedangkan f1-score untuk melihat keseimbangan presisi dengan recall. Pada hasil kelas yang diklasifikasi yang menunjukkan f1-score terbaik adalah kelas adult yang memiliki nilai presisi dan recall yang tinggi.

3.5. Optimasi Model dan Fine Tuning

3.5.1. Optimasi Hyperparameter

Untuk mendapatkan hasil yang maksimal, maka dilakukan optimasi pada hyperparameter algoritma Random Forest sebanyak lima kali percobaan. Berikut rangkuman hasil dari percobaan beberapa parameter pada algoritma Random Forest untuk menemukan model yang paling sesuai.

Model RF	n_estimator	max_depth	random_state	Akurasi Data Train	Akurasi Data Test
Model RF 1	150	10	100	91.46%	82.66%
Model RF 2	300	20	42	1.0%	81.92%
Model RF 3	450	65	50	1.0%	81.92%
Model RF 4	500	90	65	1.0%	82.25%
Model RF 5	250	25	25	1.0%	81.67%

Dari hasil percobaan parameter diatas nilai akurasi pada data test yang paling tinggi ada pada Model RF 1 dengan menggunakan n_estimator 150, max_dept 10, dan random_state 100. Dengan Model RF 5 memiliki nilai akurasi data test paling rendah sebesar 81.67%

3.5.2. Fine Tunning Evaluasi Model

Selanjutnya dilakukan Fine Tuning dengan Model KNN, untuk menemukan model terbaik untuk mengatasi kasus klasifikasi usia abalone dengan melihat dari ukuran fisiknya. Pada fine tunning ini dilakukan optimasi hyperparameter sebanyak lima percobaan. Berikut hasil dari percobaan pada tiap parameternya pada tabel dibawah :

Model KNN	n_neighbors	metric	algorithm	Akurasi Data Train	Akurasi Data Test
Model KNN 1	16	minkowski	auto	84.28%	80.59%
Model KNN 2	25	euclidean	ball_tree	84.21%	80.43%
Model KNN 3	10	manhattan	kd_tree	85.17%	80.76%
Model KNN 4	35	manhattan	brute	83.71%	80.34%
Model KNN 5	30	euclidean	kd_tree	83.71%	80.43%

Hasil akurasi Data test paling tinggi ada pada Model KNN 3 yaitu 80.76% dengan menggunakan n neighbors 10, metric manhattan, dan algoritma kd_tree. Namun nilai akurasi ini masih lebih kecil daripada Model RF 1 yang memiliki nilai akurasi data testing sebesar 82.66%. Oleh karena itu, Model RF 1 menjadi model yang terbaik untuk kasus ini.

Berikut Perbandingan dari kedua Model diatas berdasarkan Nilai Akurasinya:

Model Fine Tuning	Akurasi Data Train	Akurasi Data Test	Model RF	Akurasi Data Train	Akurasi Data Test
Model KNN 1	84.28%	80.59%	Model RF 1	91.46%	82.66%
Model KNN 2	84.21%	80.43%	Model RF 2	1.0%	81.92%
Model KNN 3	85.17%	80.76%	Model RF 3	1.0%	81.92%
Model KNN 4	83.71%	80.34%	Model RF 4	1.0%	82.25%
Model KNN 5	83.71%	80.43%	Model RF 5	1.0%	81.67%

Dapat dilihat bahwa model Random Forest memberikan rata-rata nilai akurasi lebih tinggi daripada model KNN, sehinggan model yang digunakan yaitu model Random Forest 1 karena memiliki nilai akurasi pada data testingnya paling tinggi diantara model lainnya. Model RF1 digunakan dikarenakan model RF lainnya memiliki nilai akurasi lebih kecil namun memerlukan parameter yang lebih besar sehingga membutuhkan resource yang lebih tinggi pula. Oleh karena itu, model RF 1 paling ideal dalam kasus ini.

3.6. Interprestasi dan Visualisasi Hasil

3.6.1. Terjemahan Hasil Model

Karena model RF 1 memiliki nilai akurasi yang baik, maka pada penyelesaian tugas ini akan menggunakan metode Random Forest dalam melakukan klasifikasi pada umur abalone dari pengukuran fisiknya menajdi tiga kelas yait Old, Adult, dan Young. Disini model Random Forest menggunakan n _estimator sebesar 150, dengan max_depth 10 dan random_state 100 menghasilkan akurasi training sebesar 91.465 dengan akurasi testing 82.669 seperti gambar dibawah.

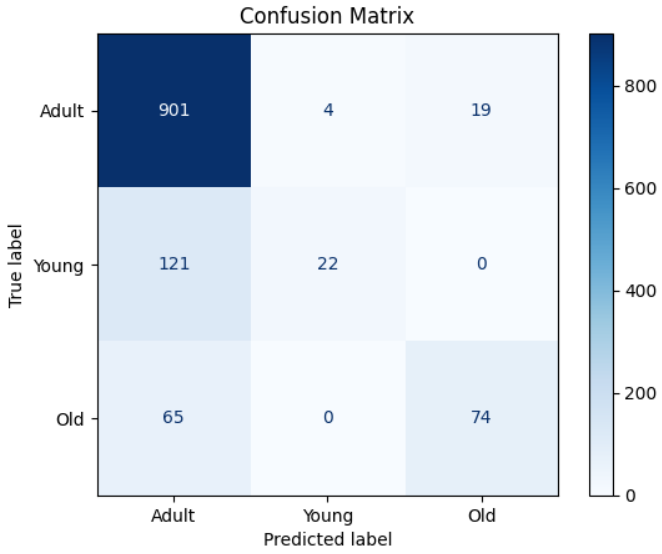
```
Random Forest Training Accuracy:  0.914651493598862
Random Forest Test Accuracy:  0.8266998341625207

Classification Report:
              precision    recall  f1-score   support

       Adult       0.83      0.98      0.90       924
         Old       0.85      0.15      0.26       143
       Young       0.80      0.53      0.64       139

    accuracy                           0.83      1206
   macro avg       0.82      0.55      0.60      1206
weighted avg       0.83      0.83      0.79      1206
```

Berikut hasil Confusion Matrix hasil dari pengklasifikasian usia abalone dengan algoritma Random Forest yang telah dibuat :



Berikut penjelasan dari confusion matriks diatas :

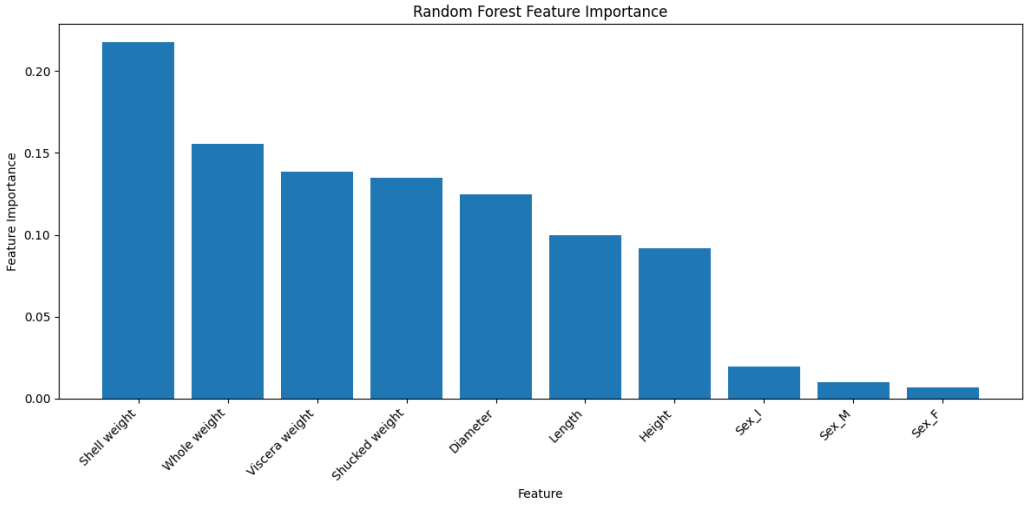
a. Untuk kelas Adult:
   - TP (True Positive) = 901, artinya 901 data tergolong dalam kelas Adult dan diprediksi benar oleh model sebagai Adult.
   - FN (False Negative) = 4, artinya 4 data tergolong dalam kelas Adult tetapi salah diprediksi sebagai Young.
   - FP (False Positive) = 19, artinya 19 data yang sebenarnya bukan tergolong dalam kelas Adult tetapi diprediksi sebagai Adult oleh model.
   - TN (True Negative): Tidak diberikan dalam confusion matrix karena tidak relevan untuk kelas yang bersangkutan.

b. Untuk kelas Young:
   - TP = 121, artinya 121 data tergolong dalam kelas Young dan diprediksi benar oleh model sebagai Young.
   - FN = 22, artinya 22 data tergolong dalam kelas Young tetapi salah diprediksi sebagai Adult.
   - FP = 0, artinya 0 data yang sebenarnya bukan tergolong dalam kelas Young tetapi diprediksi sebagai Young oleh model.

c. Untuk kelas Old:

- TP = 65, artinya 65 data tergolong dalam kelas Old dan diprediksi benar oleh model sebagai Old.
- FN = 0, artinya 0 data tergolong dalam kelas Old tetapi salah diprediksi sebagai Young.
- FP = 74, artinya 74 data yang sebenarnya bukan tergolong dalam kelas Old tetapi diprediksi sebagai Old oleh model.

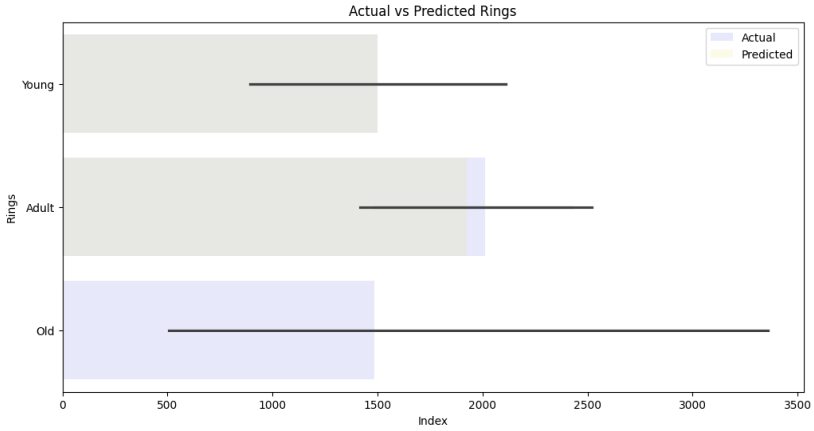
Berikut Fitur yang memiliki peranan penting dalam model yang telah dibuat untuk mengklasifikasikan umur abalone berdasaran pengkuran fisiknya :



Dapat dilihat bahwa fitur yang memiliki peranan penting dalam pengklasifikan umur abalone menggunakan pengukuran fisik adalah fitur Shell Weight, fitur weight atau berat menjadi fitur penting diikutin fitur lainnya. Dimana fitur Shell weight atau berat cangkang kerang menunjukkan semakin besar ukurannya, dan menjadi fitur yang paling berperan penting pada penentuan usia abalone.

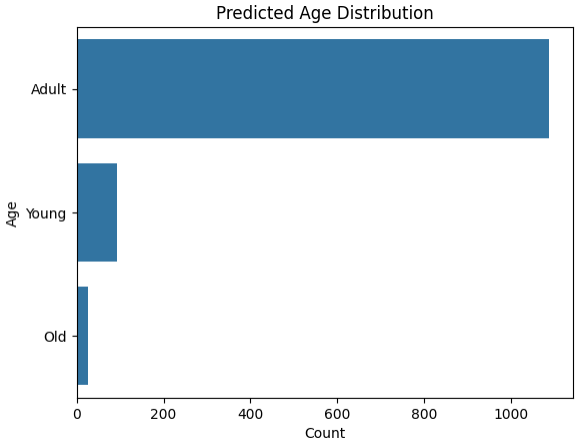
3.6.2. Visualisasi Hasil Model

Berikut adalah visualisasi dari hasil klasifikasi menggunakan model Random Forest dilakukan perbandingan dengan dataset yang ada, ditunjukkan pada gambar dibawah ini.

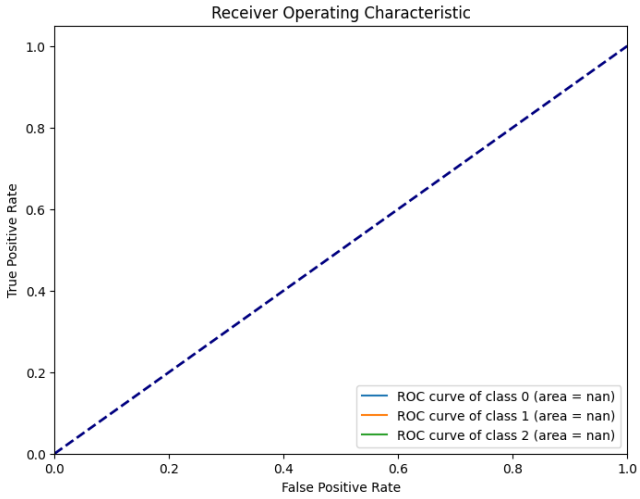


Dapat dilihat dari barchart perbandingan antara hasil klasifikasi dan data asli terdapat pada kelas Young berhasil diklasifikasikan dengan baik dan tidak telihat perbedaandaan dari hasil klasifikasi dengan data aslinya. Pada kelas Adult terdapat data asli yang tidak terklasifikasi, namun memiliki hasil klasifikasi terbanyak yang benar. Dan kelas Old hasil klasifikasinya terlihat belum bisa mengklasifikasikan data Old dengan benar, karena kelas Old memiliki data yang lebih sedikit dibandingkan dengan kedua kelas usia lainnya.

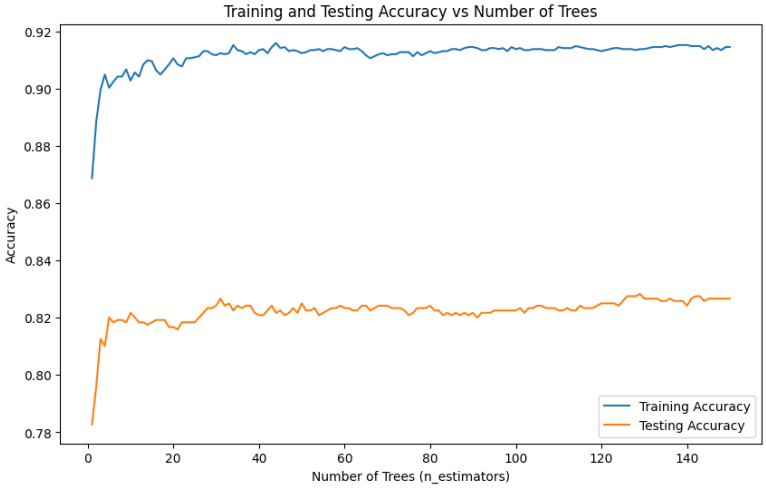
Kemudian, dilakukan perhitungan pesebaran hasil klasifikasi usia abalone dengan model Random Forest yang telah dibuat didapatkan seperti gamabr dibawah.



Dari hasil klasifikasi dari model yang telah dibuat pesebaran data usia pada abalone didominasi oleh kelas Adult, yang kemudian ad akelas Young, dan kelas yang paling kecil adalah Old.



Pada kurva ROC diatas ketiga kelas hasil klasifikasi memiliki kurva yang sama, dimana semakin tinggi true positive makin meningkat pula false negativenya. Yang menandakan bahwa model Random Forest untuk melakukan klasifikasi usia abalone dengan ukuran fisikanya yang telah dibuat ini memiliki tingkat kesalahan yang meningkat dalam mengklasifikasikan contoh negatif sebagai positif (FPR meningkat), sementara juga memiliki tingkat keberhasilan yang meningkat dalam mengklasifikasikan contoh positif dengan benar (TPR meningkat).



Grafik nilai akurasi pada training dan testing model diatas menunjukkan semakin meningkatnya n_estimator maka semakin meningkat pula nilai akurasinya, baik pada data training maupun testing, grafik akurasi menunjukkan peningkatan yang stabil seiring berjalannya waktu, ini menunjukkan bahwa model terus belajar dan meningkatkan kemampuannya untuk mengklasifikasikan tiap fitur dengan benar.

Pada grafik loss pada training dan testing model yang telah dibuat menunjukkan bahwa tiap naiknya n_estimator atau jumlah pohon pada algoritma Random Forest maka akan semakin menurun pula nilai loss nya. Dimana grafik loss yang menunjukkan penurunan yang stabil seiring berjalannya waktu, menandakan bahwa model yang telah dibuat secara bertahap mengurangi kesalahannya dan semakin mendekati kemampuan optimalnya.

BAB IV
PENUTUP

4.1. Kesimpulan

Dari percobaan yang telah dilakukan untuk melakukan klasifikasi usia abalone dari ukuran fisiknya, dilakukan fine tunning dengan dua algoritma yaitu Random Forest dan KNN. Dimana setiap algoritma dilakukan lima kali percobaan hyperparameter untuk mendapatkan model yang paling efektif untuk kasus ini. Dimana ketika dibandingkan antara kedua algoritma yang digunakan, algoritma Random Forest lebih unggul dalam melakukan klasifikasi daripada KNN. Dimana nilai rata-rata akurasi pada algortima Random Forest lebih tinggi dibandingkan akurasi KNN. Dimana nilai akurasi paling tinggi didapatkan pada model Random Forest 1, yang menggunakan *n_estimator* 150 dengan max_depth 50 menghasilkan akurasi 82.66%, dan nilai akurasi paling kecil ada pada algoritma KNN 4 dengan *n_estimator* 35, matrik Manhattan, dan algoritma brute menghasilkan nilai akurasi 80.34% Pengklasifikasian usia abalone dari pengukuran fisik sangat dipengaruhi dari seluruh variable ukuran fisiknya, ukuran fisik seperti panjang, diameter, tinggi, serta bobot total, bobot daging, bobot viscera, dan bobot cangkang memiliki hubungan yang signifikan dengan usia abalone.

4.2. Saran

Hasil penelitian ini meningkatkan pemahaman kita tentang faktor-faktor yang mempengaruhi pertumbuhan abalone berdasarkan ukuran fisiknya. Rekomendasi untuk penelitian selanjutnya adalah penggunaan metode pembelajaran mesin lainnya atau kombinasi metode statistik tradisional untuk meningkatkan akurasi klasifikasi usia abalone. Selain itu, penelitian lebih lanjut juga dapat dilakukan untuk mengeksplorasi faktor-faktor lingkungan lainnya yang dapat memengaruhi pertumbuhan dan usia abalone.