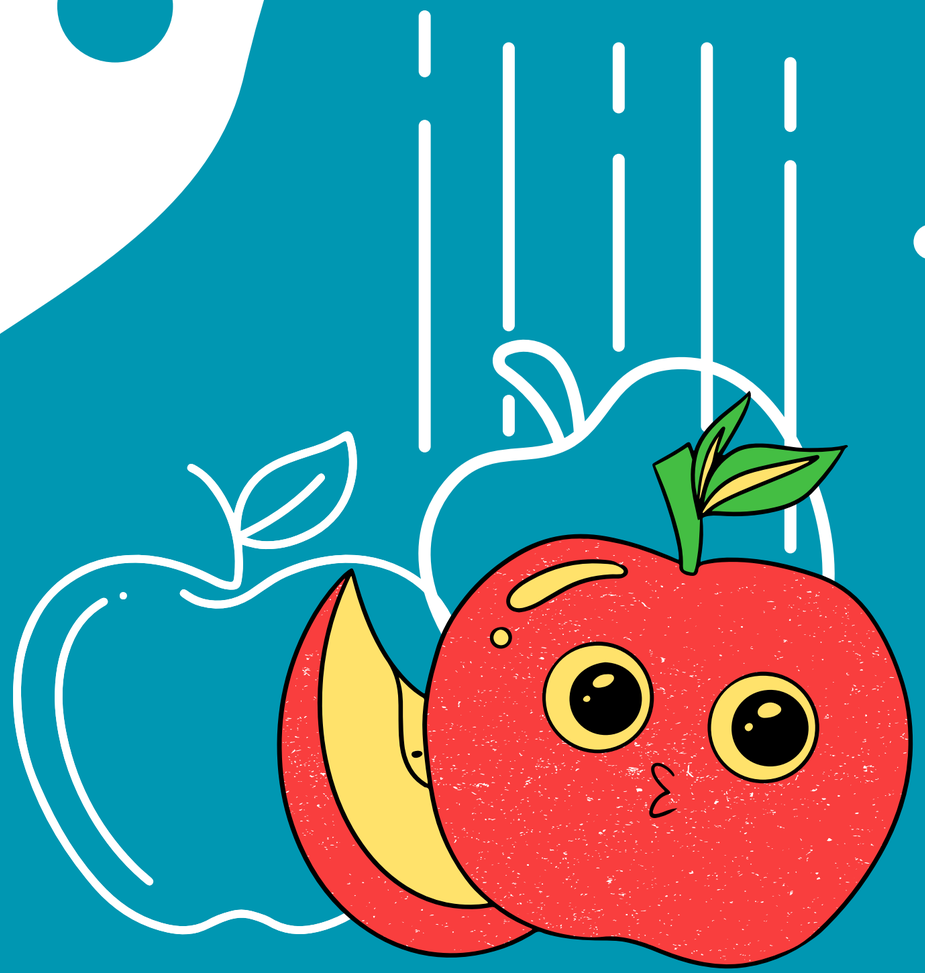
KLASIFIKASI QUALITY APEL

By : Noeril Agian Septa Dinata



#Data Science #Portfolio #Data Analyst

Quality apel??

Tujuan dari penelitian ini adalah mengetahui apa saja yang paling berpengaruh terhadap klasifikasi quality (good and bad) pada buah apel

(4001, 9)

Data Preparation

Berikut merupakakn data sampel dari data klasifikasi buah apel

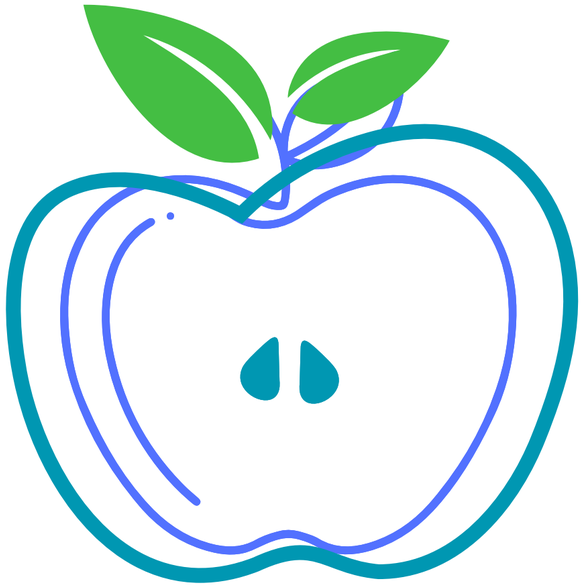
Sampel Data :

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
0	0.0	-3.970049	-2.512336	5.346330	-1.012009	1.844900	0.329840	-0.491590483	good
1	1.0	-1.195217	-2.839257	3.664059	1.588232	0.853286	0.867530	-0.722809367	good
2	2.0	-0.292024	-1.351282	-1.738429	-0.342616	2.838636	-0.038033	2.621636473	bad
3	3.0	-0.657196	-2.271627	1.324874	-0.097875	3.637970	-3.413761	0.790723217	good
4	4.0	1.364217	-1.296612	-0.384658	-0.553006	3.030874	-1.303849	0.501984036	good

Total data sebanyak (baris, kolom) : 4001, 9

A_id: *Unique identifier for each fruit*
Size: *Size of the fruit*
Weight: *Weight of the fruit*
Sweetness: *Degree of sweetness of the fruit*
Crunchiness: *Texture indicating the crunchiness of the fruit*
Juiciness: *Level of juiciness of the fruit*
Ripeness: *Stage of ripeness of the fruit*
Acidity: *Acidity level of the fruit*
Quality: *Overall quality of the fruit*

Link data : https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality



Penggalian Masalah pada Data

MISSING VALUE

Terdapat missing sebanyak 1 baris, solusi yang dilakukan yaitu remove data kosong/ null karena hanya ada 1 data yang kosong yaitu terletak pada baris 4000.

TYPE DATA

Pada type data tidak ditemukannya masalah, semua type data benar

DUPLICATE

Tidak ditemukannya data duplicate

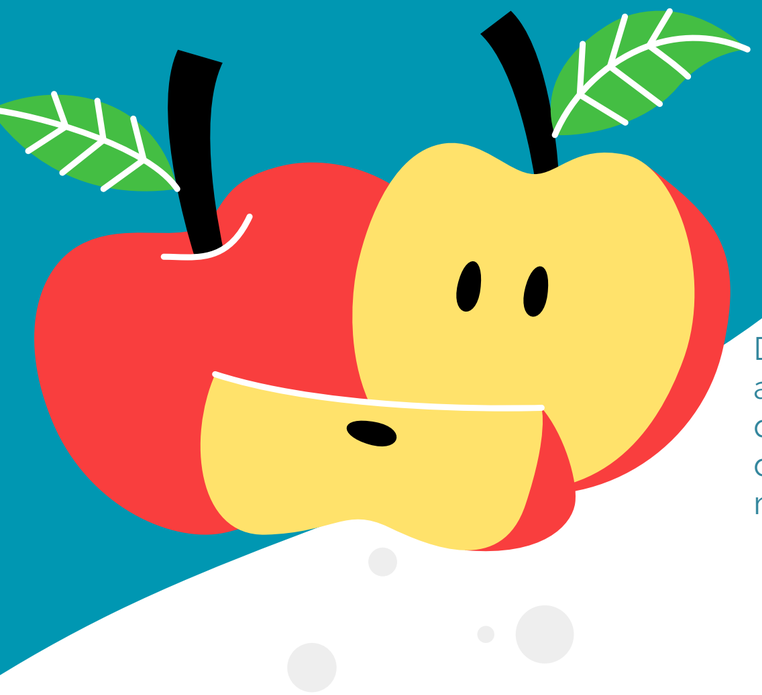
sampel data yang missing :

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
4000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Created_by_Nidula_Elgiriyewithana	NaN

Exploratory data analysis and Modelling

Pada tahapan EDA kita dapat mengetahui persebaran dan informasi yang ada dalam data kemudian tahapan pemodelan ini kita akan dibantu oleh machine untuk memprediksi. pemodelan ini akan menghasilkan nilai dari suatu algoritma yang nantinya akan membantu kita menyelesaikan masalah klasifikasi apel.

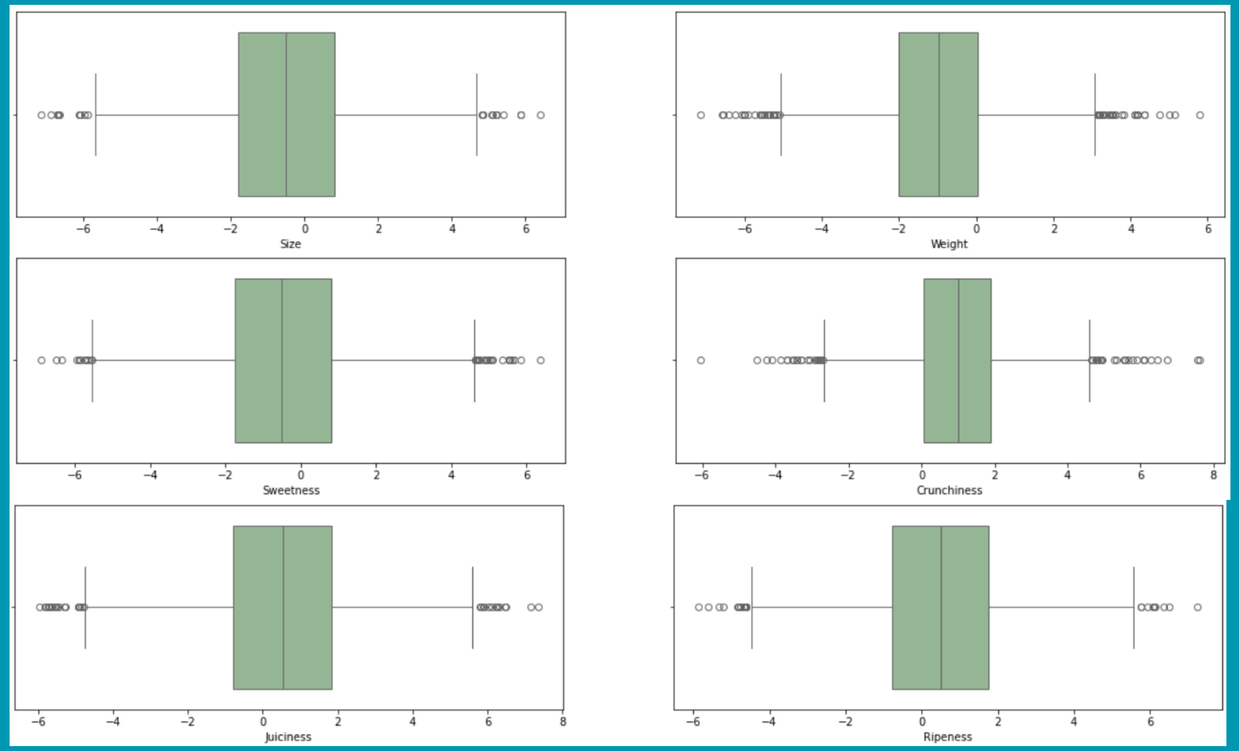
Exploratory data analysis

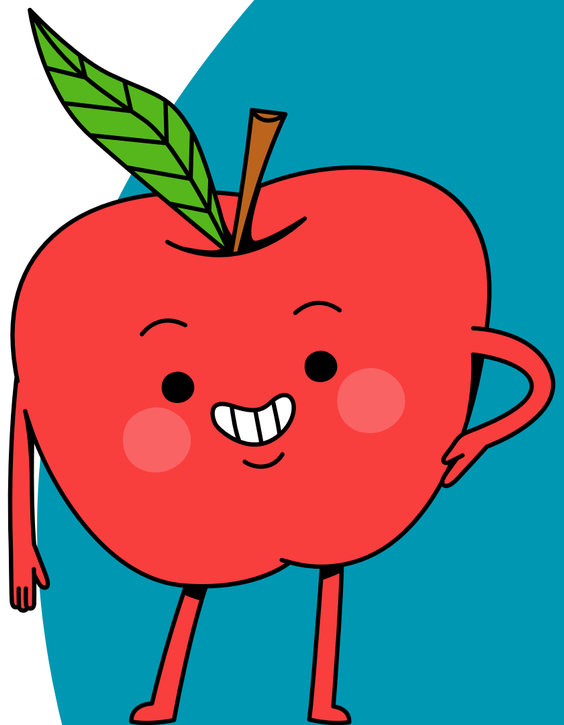


Visualisasi Data:

Dari visualisaasi ini dapat diketahui persebaran data yang ada dalam data klasifikasi apel dan juga diketahui nilai data quality apel cukup seimbang antara good dan bad. Namun dari visualisasi tersebut dapat dililhat bahwasannya akan memiliki nilai outlier.

Outlier





Pada data terdapat outlier, solusi yang saya lakukan adalah membiarkannya karena outlier merupakan data yang penting dan dapat membantu meningkatkan hasil prediksi kita, namun dalam artian outlier yang kita temui ini merupakan penilaian dari sebuah pengamatan sebuah apel jadi outlier masih aman, berbeda dengan outlier terhadap pembelian dll, hal tersebut harus di atasai karena akan berdampak negatif apabila tidak di atasi, Sebenarnya banyak solusi untuk mengatasi outlier contohnya menghapus, mengubah, mencari info lagi tentang data yang outlier bisa menanyakan kepada penyedia data atau bisa berkaloborasi dengan data enginer..

Data Processing

Melakukan Label Encoding

Tujuan untuk mengubah kategorical kedalam bentuk numerik

Algoritma Random Forest dan Decision Tree

Machine learning digunakan untuk mengklasifikasi

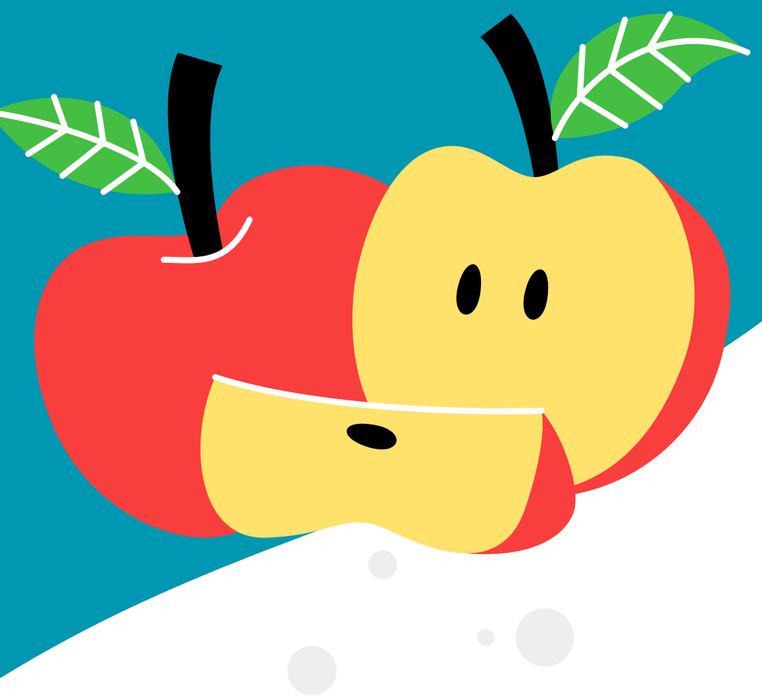
Menampilkan hasil Akurasi, Presisi, Recall.

Membandingkan Algoritma terbaik

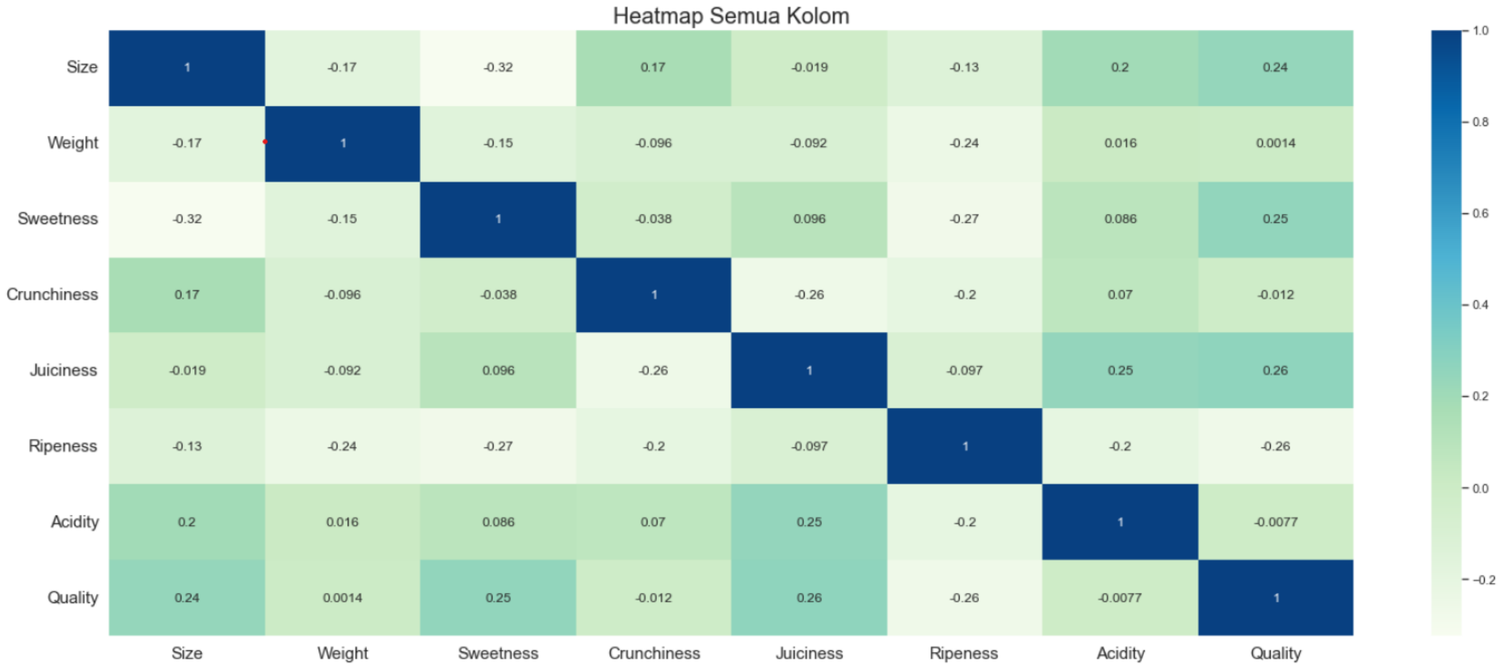
Feature Importance

Melakukan prediksi kolom yang berpengaruh terhadap quality berdasarkan algoritma yang memiliki nilai akurasi tertinggi.

Visualisasi Heatmap



Visualisasi dengan Heatmap:



visualisasi ini bertujuan untuk mengetahui pendekatan atau hubungan dari setiap kolomnya, dari visualisasi ini dapat kita lihat bahwa ada beberapa kolom yang saling berhunungan diantaranya adalah **Quality dengan Juiciness, Quality dengan Sweetness, dan Quality dengan Size**, jadi dapat diketahui atau dapat diprediksi awal yang mempengaruhi quality dari klasifikasi apet yaitu **Juiciness, Sweetness, dan Size,** ini hanya prediksi awal untuk mengetahui kebenarannya nanti akan dibantu oleh algoritma machine learning.,

Hasil dari Algoritma

Dari salah satu algoritma akan dibandingkan dan apabila nilai terdapat lebih tinggi dari algoritma lainnya maka algoritma tersebut yang akan belanjut ke Feature Importance

Berikut sampel data yang sudah melalui tahap encoding

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
0	0.0	-3.970049	-2.512336	5.346330	-1.012009	1.844900	0.329840	-0.491590483	1
1	1.0	-1.195217	-2.839257	3.664059	1.588232	0.853286	0.867530	-0.722809367	1
2	2.0	-0.292024	-1.351282	-1.738429	-0.342616	2.838636	-0.038033	2.621636473	0
3	3.0	-0.657196	-2.271627	1.324874	-0.097875	3.637970	-3.413761	0.790723217	1

Berikut hasil algoritma Random Forest & Decision Tree

```
Decision Tree Classifier:
Accuracy: 0.81
Confusion Matrix:
 [[324  77]
 [ 75 324]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.81      0.81       401
           1       0.81      0.81      0.81       399

    accuracy                           0.81       800
   macro avg       0.81      0.81      0.81       800
weighted avg       0.81      0.81      0.81       800
```

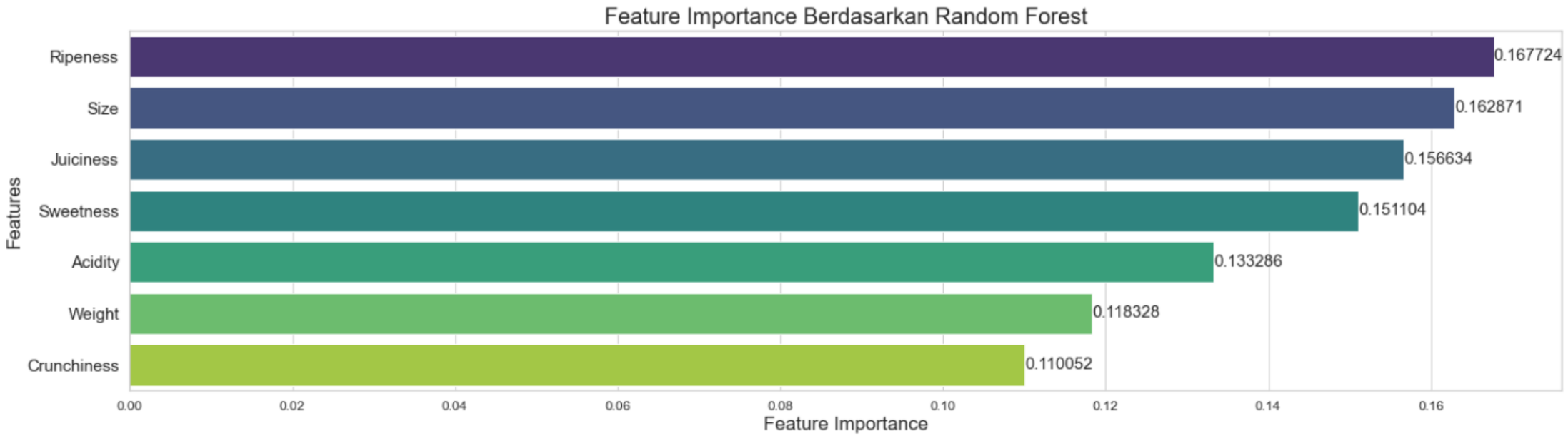
```
Random Forest Classifier:
Accuracy: 0.9075
Confusion Matrix:
 [[364  37]
 [ 37 362]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       401
           1       0.91      0.91      0.91       399

    accuracy                           0.91       800
   macro avg       0.91      0.91      0.91       800
weighted avg       0.91      0.91      0.91       800
```

Dapat kita lihat bahwa Algoritma Random Forest lebih unggul Maka Algoritma tersebut lanjut ke step selanjutnya yaitu feature importance untuk memprediksi kolom kolom yang mempengaruhi good dan bad pada data klasifikasi apel.

Fiture Importance



Dapat kita lihat yang paling mempengaruhi klasifikasi apel good atau bad adalah Ripeness, Size, Juiciness, Sweetness

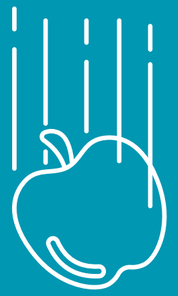
Cek kebenaran menggunakan data max dan min :

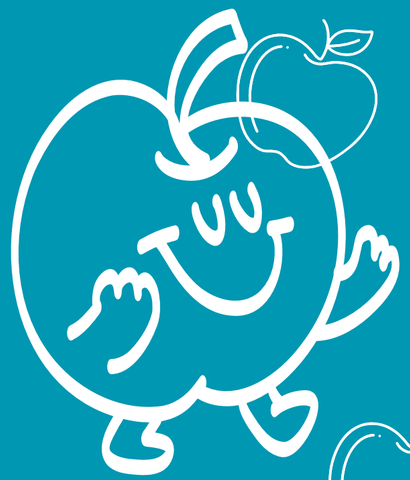
Nilai Max :

	Ripeness	Size	Juiciness	Sweetness	Quality
1381	3.995602	6.406367	0.999640	-4.164118	1
2502	7.237837	-0.921677	0.447844	-4.685317	0
2691	-0.336443	-4.122996	2.835063	6.374916	1
3874	-1.342563	1.925888	7.364403	-0.127441	1

Nilai Min :

	Ripeness	Size	Juiciness	Sweetness	Quality
161	-5.864599	3.256911	-4.283278	-0.403344	0
2232	0.301017	-0.214485	-5.961897	-3.942249	0
2832	2.291800	-1.520566	0.129411	-6.894485	0
3559	4.127709	-7.151703	-0.136146	1.673138	0





Kesimpulan and Saran

Kesimpulan :

Diketahui algoritma memiliki nilai tertinggi adalah **random forest** dan berdasarkan hasil feature importance dari random fores dapat dilihat kolom yang sangat perpengaruh terhadap quality good atau bad pada sebuah klasifikasi apel adalah (Ripeness, Size, Juiciness, Sweetness)

- Apabila Ripness tinggi maka quality dari aple tersebut terprediksi **Bad**
- Jika Size, Juiciness, Swetness bernilai tinggi maka aple dikategoruikan aple **good** dan begitupun sebaliknya apabila nilai rendah maka apel akan bernilai **Bad**

Dari hasil prediksi yang ditampilkan menggunakan feature importance di atas dapat dibuktikan prediksi benar

Saran:

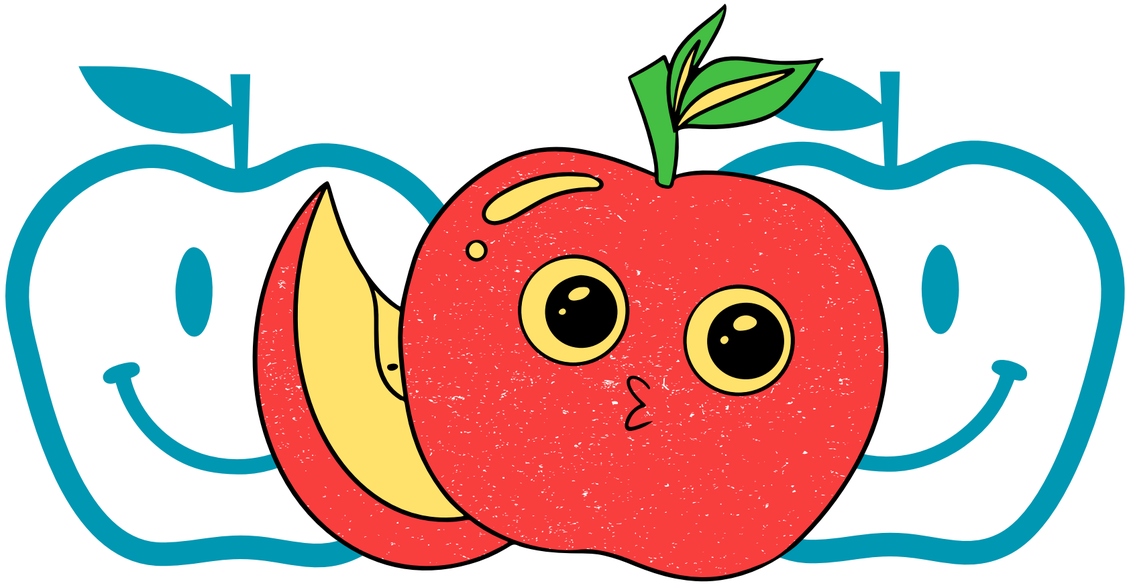
Jadi apabila ingin memilih apel dengan kualitas **GOOD** pililhlah dengan **Size besar**, rasanya **Juiciness** dan **Sweetness**, namun apabila berdasarkan nilai pilih juga dengan nilai **Ripeness rendah**

Thanks

Do you have any questions?

Apabila ada kesulitan atau ada yang perlu dibahas silahkan DM. **saya harap pnelitian ini tidak hanya sampai disini, anda bisa mencoba dengan algoritma lain**.

Terima Kasih



#Data Science