



Supervised Learning Report for Prasmulyan's Preference of Clothing

PRESENTED BY:

Jovanka Cathrynn Thomas	23102010021
Felix Budhi	23102010028
Brandon Haris	23102010051
Marchella Veranika Tobing	23102010054



Results accompanied with html markdown.

ABOUT DATA SET

Raw Data: 310 Rows (307 Prasmulyan, 3 Non - Prasmulyan) dan 21 Columns

1 PakaianKampus

2 BudgetFashion

3 BudgetKaos

4 BudgetKemeja

5 JumlahPakaianPurchased

6 OversizedSlimfit

7 Keringat

8 DosenFormal

9 JumlahKemeja

10 JumlahKaos

11 PakaianSeringPakai

12 KaosDiKelas

13 TerangGelap

DATA CLEANING

Checking Missing Values

```
data.isnull().sum()
Apakah Anda mahasiswa Prasetya Mulya?          0
Gender                                              3
Usia                                               3
Fakultas                                           3
Jurusan                                           3
Angkatan                                           3
Domisili                                           3
Ketika Anda pergi ke kampus, apakah anda lebih suka memakai kaos atau kemeja?    3
Berapa rata-rata budget yang dikeluarkan untuk membeli pakaian per bulan?      3
Berapa rata-rata budget yang dikeluarkan untuk membeli kaos per bulan?        3
Berapa rata-rata budget yang dikeluarkan untuk membeli kemeja per bulan?       3
Berapa jumlah rata-rata pembelian pakaian dalam sebulan?                      3
Apakah Anda merupakan orang yang memperhatikan penampilan?                  3
Manakah yang menurut Anda lebih nyaman dipakai?                            3
Apakah Anda mudah berkeringat?                                         3
Apakah Anda memiliki peraturan berpakaian formal dari dosen kelas?           3
Berapa jumlah kemeja yang Anda miliki?                                3
```

Clear Non-Prasmulyan Data

```
data = data.drop(labels=[13,89,122], axis=0)
data.Prasmulyan.unique()
array(['Iya'], dtype=object)
```

Rename Columns

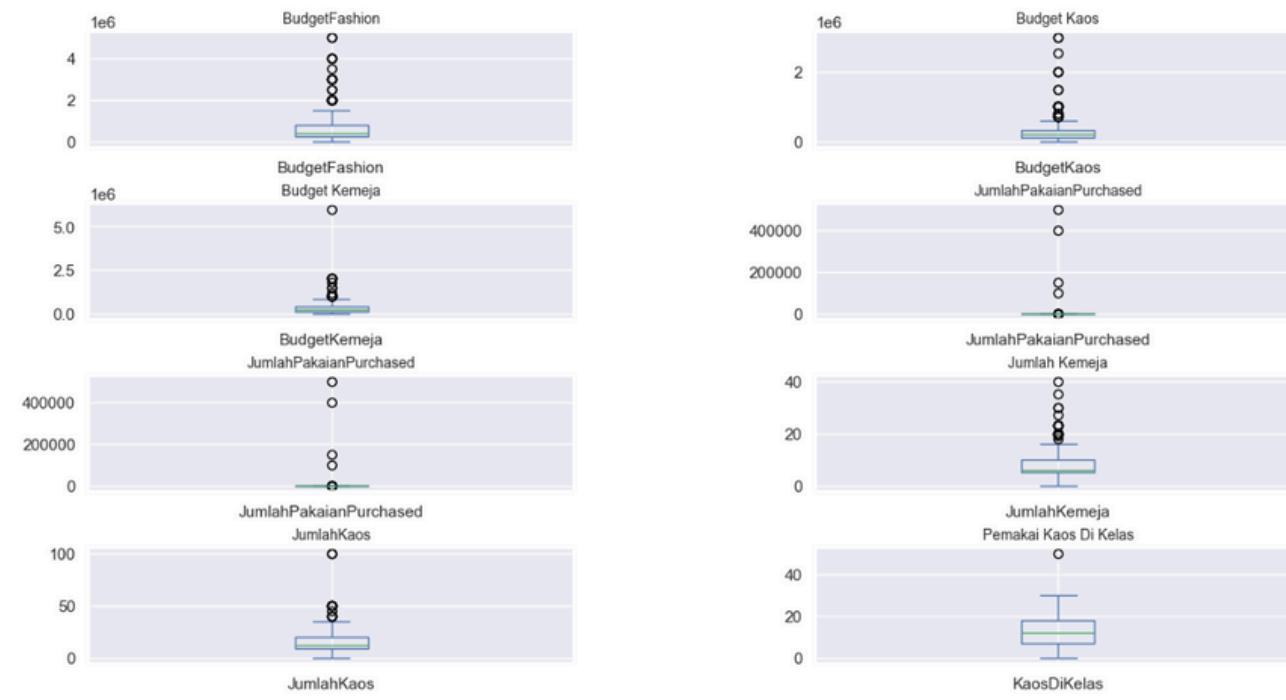
```
df.dtypes
PakaianKampus          int64
BudgetFashion           int64
BudgetKaos              int64
BudgetKemeja             int64
JumlahPakaianPurchased int64
JAIM                   int64
OversizeSlimfit         int64
Keringat               int64
DosenFormal             int64
JumlahKemeja            int64
JumlahKaos              int64
PakaianSeringPakai      int64
KaosDiKelas             int64
TerangGelap              int64
dtype: object
```

Clean Data from Invalid Data

```
invaliddata1 = df[df['JumlahKemeja'].str.contains('invalid')]
print(invaliddata1)
   PakaianKampus BudgetFashion BudgetKaos BudgetKemeja JumlahPakaianPurchased \
33     Kaos          125000      50000       100000          2
   JAIM          OversizeSlimfit Keringat DosenFormal JumlahKemeja JumlahKaos \
33   Iya            Slimfit      Iya      Iya      invalid       10
   PakaianSeringPakai KaosDiKelas TerangGelap
33        Kaos            10        Gelap
```

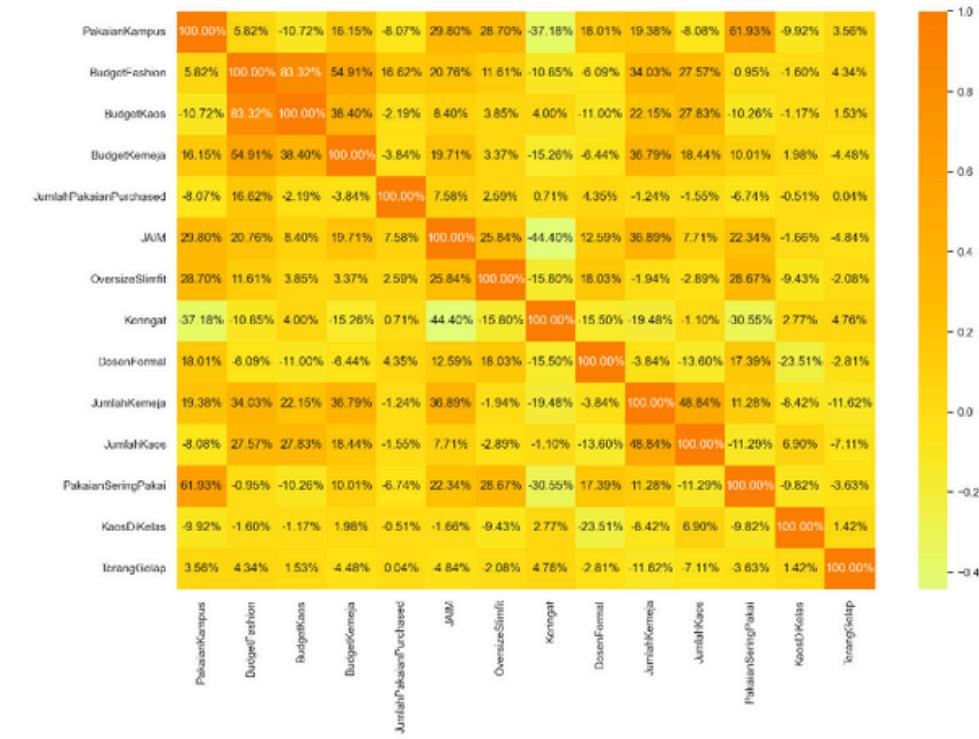
Exploratory Data Analysis (EDA)

Boxplot



Dari Boxplot diatas kita dapat melihat banyak sekali outliers yang nampak pada sebagian besar data. Oleh karena itu, kita akan melakukan pembersihan outlier.

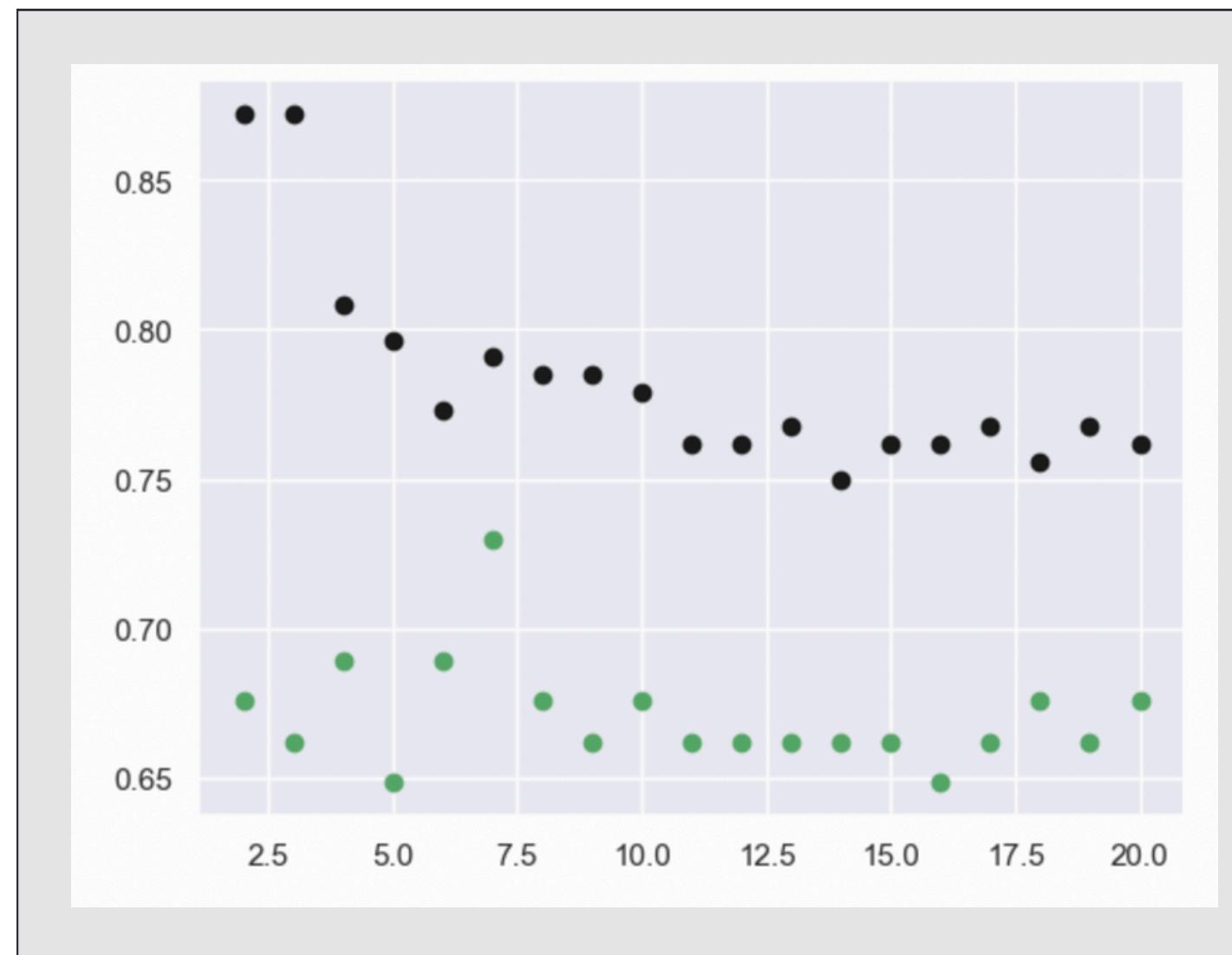
Heatmap



Dari Heatmap diatas, ditunjukkan bahwa variabel yang paling berpengaruh signifikan terhadap variabel Y adalah Pakaian Bepergian, dimana Pakaian Bepergian dapat mencerminkan korelasi positif sebesar 61.93% terhadap Pakaian yang digunakan saat mahasiswa pergi ke kampus.

K-Nearest Neighbor (KNN)

Berikut hasil dari KNN :



Result:

TIDAK AKURAT

Reason:

Generalisasi data yang teriterasi secara berlebihan.

DECISION TREE

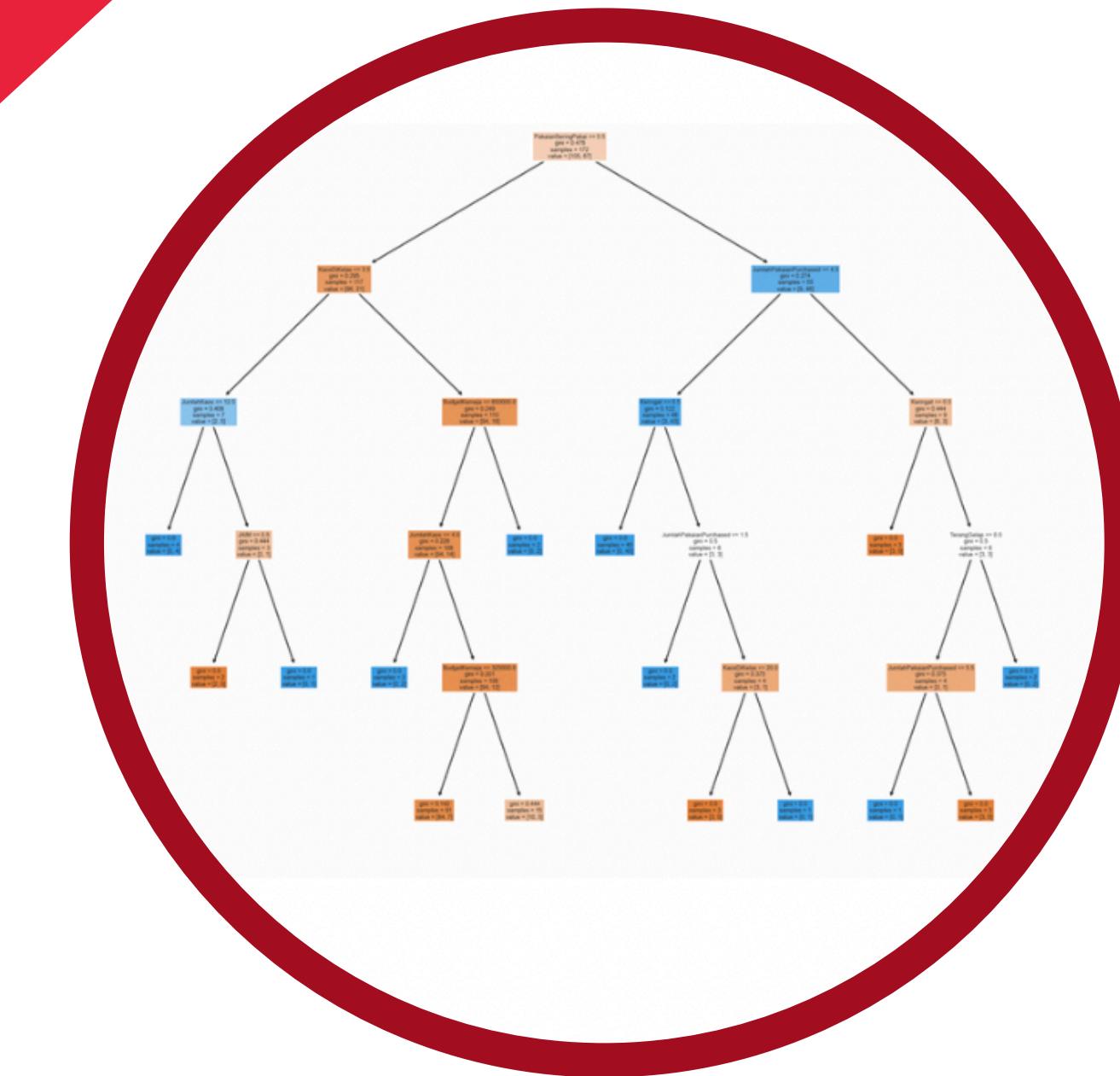
Berikut hasil dari Decision Tree :

Result :

CUKUP AKURAT (93.02%)

Details:

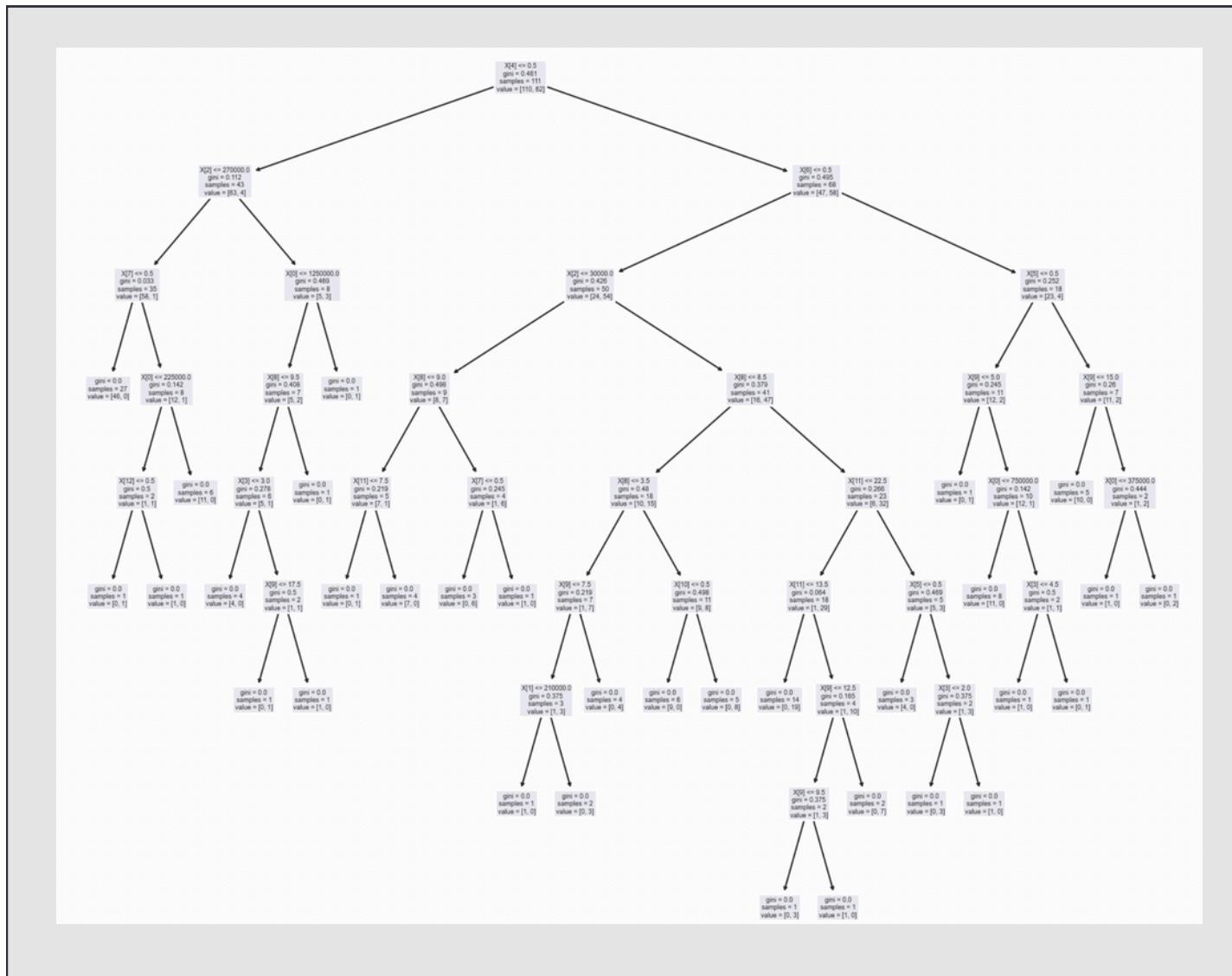
- Terdapat 5 faktor yang sangat berpengaruh.



	Metrics	Training Data	Testing Data
0	Accuracy	93.02%	81.08%
1	Sensitivity	82.09%	62.07%
2	Precision	100.00%	85.71%

RANDOM FOREST

Berikut hasil dari Random Forest :



Metrics Training Data Testing Data

0	Accuracy	93.88%	84.00%
1	Sensitivity	84.62%	77.78%
2	Precision	100.00%	77.78%

Result:

PALING AKURAT (93.88%)

Reason:

Hasil yang didapatkan *lebih stratified*.

FINAL MODEL EVALUATION

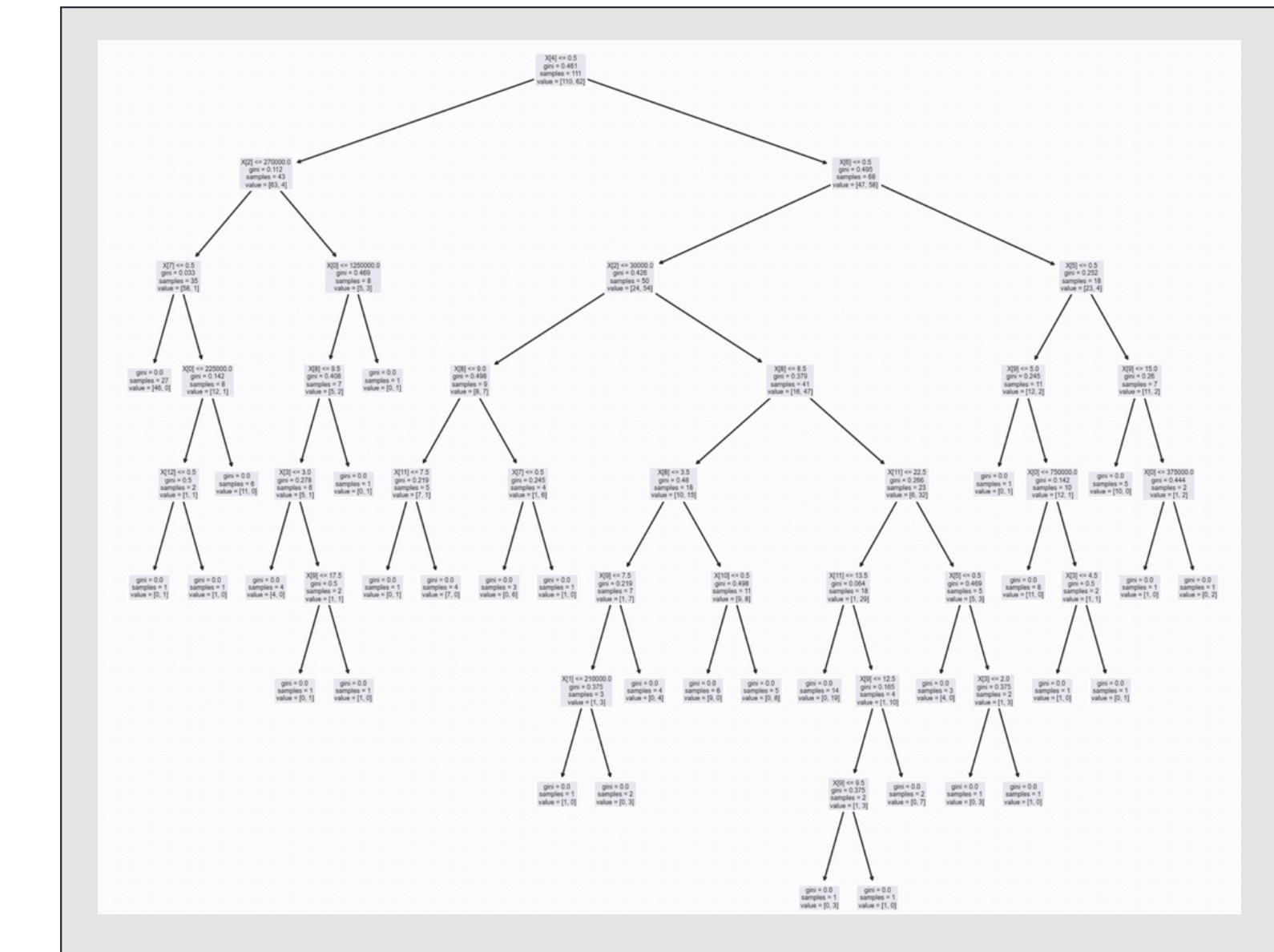
Evaluation Metrics	Decision Tree	Unoptimized Models with KNN	Optimized Models with Random Forest
Accuracy	0.9302	0.7419	0.9387
F-score	0.7804	0.6878	0.7804

Random Forest dipilih sebagai **model yang paling unggul dengan akurasi yang baik dan skor F-score pada grid cross-validation sebelumnya**. Tingkat akurasi dan skor F-score pada uji memiliki perbedaan kecil dengan skor hasil tes dari grid cross-validation, dan memiliki efek generalisasi yang baik.

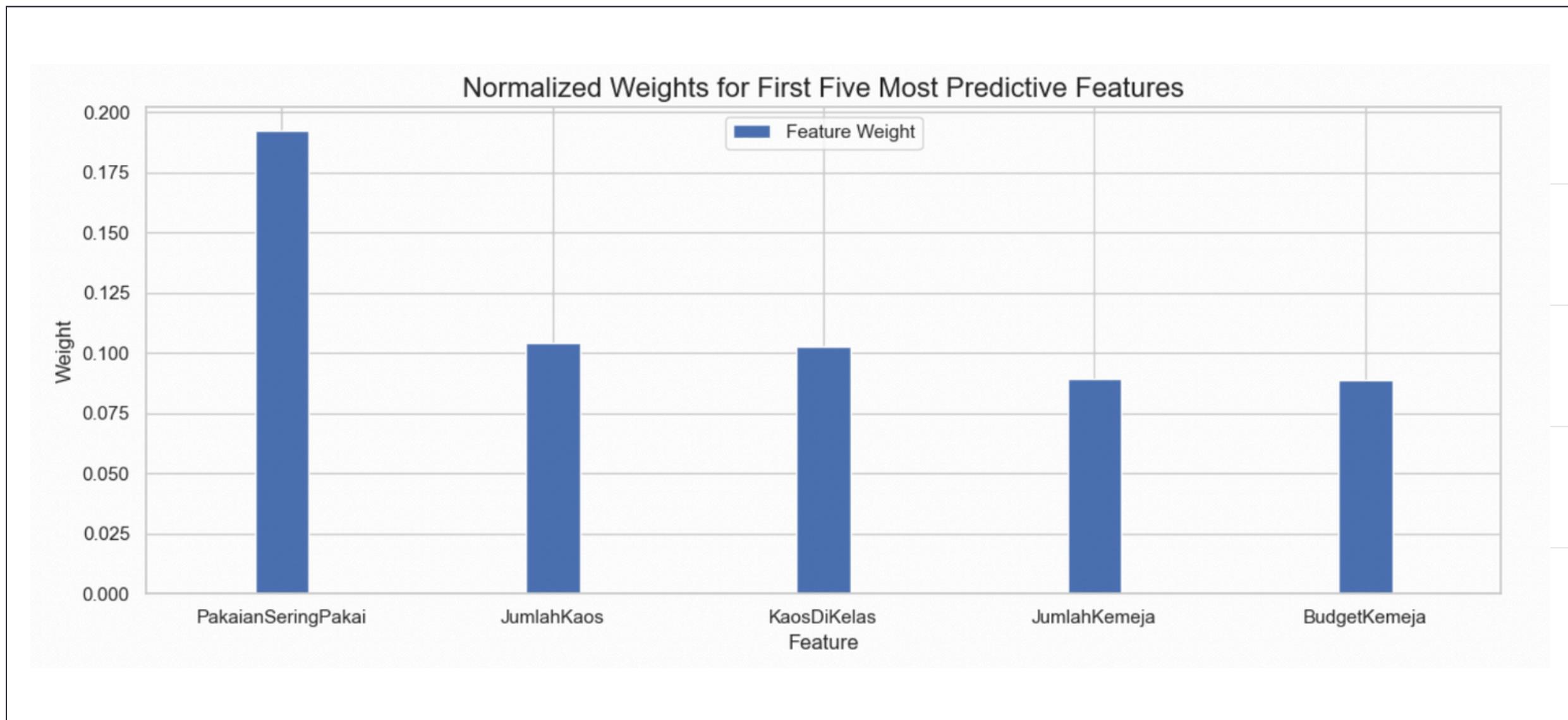
PREDICTION RESULT

Hasil Prediksi

Berdasarkan model **Random Forest**, data yang dimodelkan memprediksi mahasiswa lebih memilih menggunakan **Kaos** daripada kemeja. Hasil prediksi menunjukkan **93% (288/310)** mahasiswa cenderung menggunakan kaos.



MOST PREDICTIVE FEATURES



- 1 PakaianSeringPakai
- 2 JumlahKaos
- 3 KaosDiKelas
- 4 JumlahKemeja
- 5 BudgetKemeja

CONCLUSION

FINAL MODEL EVALUATION:

TIME	<ul style="list-style-type: none">• KNN : Memerlukan waktu paling sedikit dan hasil supervised learning yang kurang akurat.• Random Forest : Memerlukan waktu cukup lama tetapi memiliki hasil supervised learning yang paling akurat dengan branching step sebanyak 99 kali.• Decision Tree : Memerlukan waktu supervised learning yang relatif lebih cepat dibandingkan random forest dan hasil yang cukup akurat. <p>Dengan demikian, RANDOM FOREST merupakan model yang memerlukan waktu untuk melakukan supervised learning paling lama.</p>
ACCURACY & F-1 SCORE	<ul style="list-style-type: none">• KNN : Memiliki classification cukup standard dimana hasil hasil classification terbagi menjadi 2 bagian cluster.• Random Forest : Memiliki hasil akurasi yang baik dengan F-score pada grid cross-validation sebelumnya.• Decision Tree : Memiliki hasil classification and decomposition sangatlah bagus tetapi dengan hasil accuracy yang kurang baik jika dibandingkan dengan random forest. <p>Dengan demikian, RANDOM FOREST merupakan model yang memiliki hasil supervised learning paling akurat.</p>

Random Forest dipilih sebagai model yang paling akurat dengan akurasi yang baik dan skor F-score pada grid cross-validation sebelumnya meskipun memerlukan waktu lebih lama.

PREDICTION RESULT:

Prasmulyan cenderung memiliki preferensi untuk menggunakan **kaos** ketika melakukan kegiatan perkuliahan.

THANK
YOU

