UTS KLASIFIKASI MODEL

RAISYA ATHAYA KAMILAH

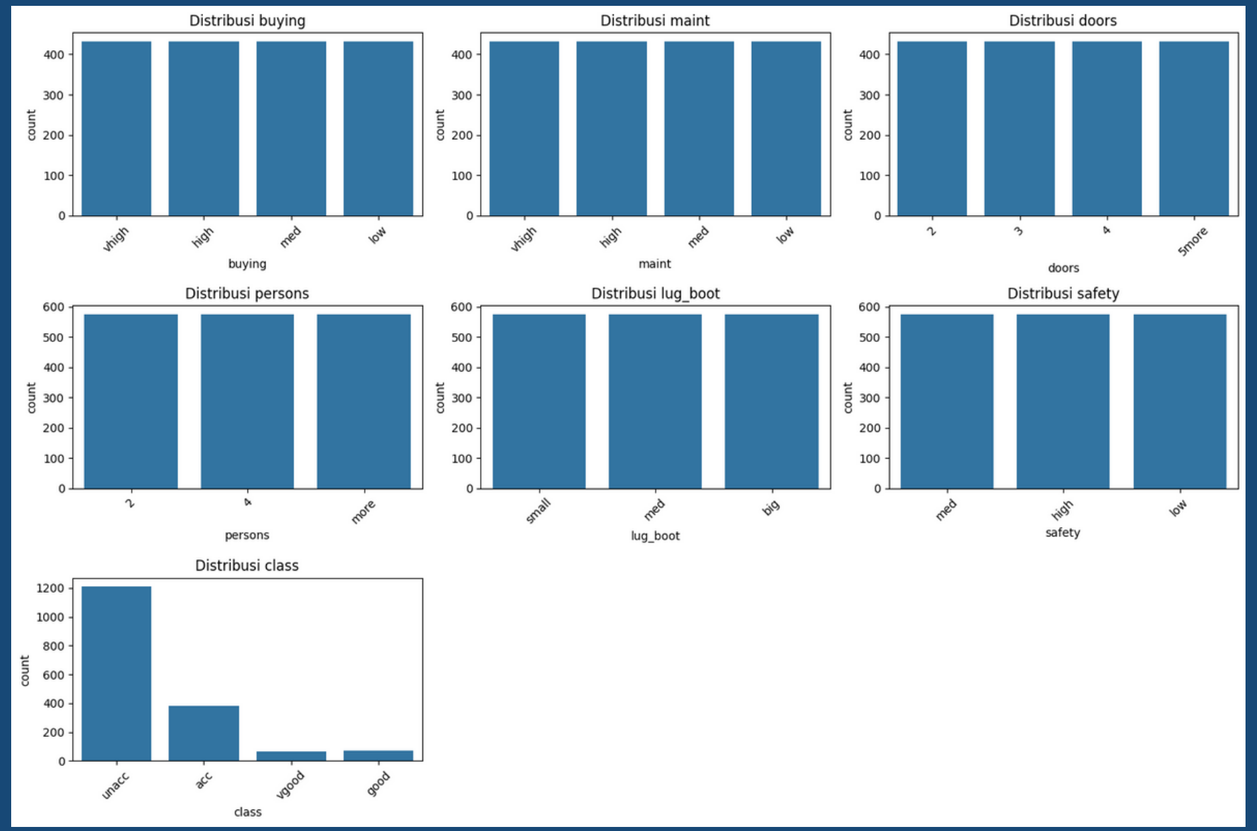
101032380253

TKX-47-01

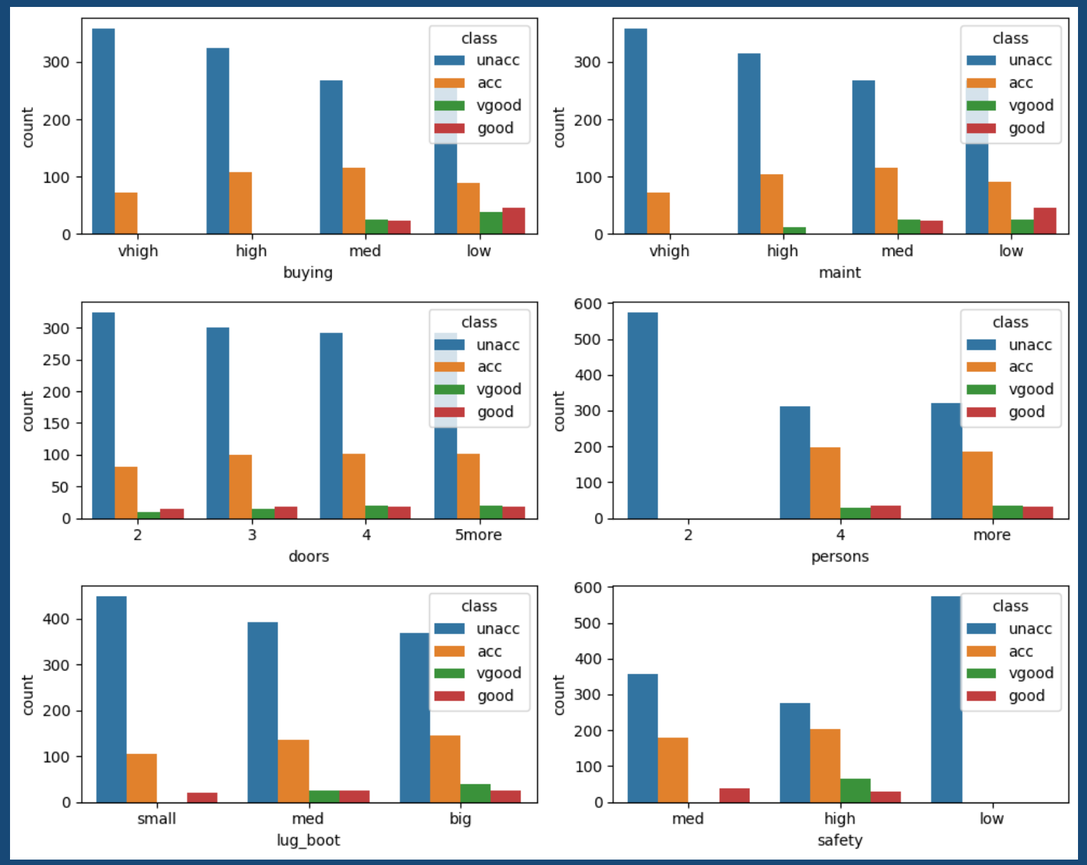
EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) adalah proses awal dalam analisis data yang bertujuan untuk mengeksplorasi dan memahami struktur, pola, hubungan, dan distribusi data sebelum melakukan analisis lebih lanjut atau pembangunan model prediktif. Tujuannya untuk memahami Karakteristik Data,Menemukan Pola dan Hubungan,Pengecekan Missing Data

VISUALIZATION DATA



Berdasarkan hasil Visualisasi data terdapat ketidakseimbangan kelas di distribusi Class



Menunjukkan kelas uacc mendominasii dihampir semua fitur yang berarti datanya sangat tidak seimbang, fitur safety dan person merupakan fitur penting karena hasil distribusinya lebih bervariasi. Fitur buying dan maint memiliki sedikit kelas good dan vgood di kategori rendah.

SEBELUM

```
class
unacc    242
acc       77
good      14
vgood     13
Name: count, dtype: int64
```

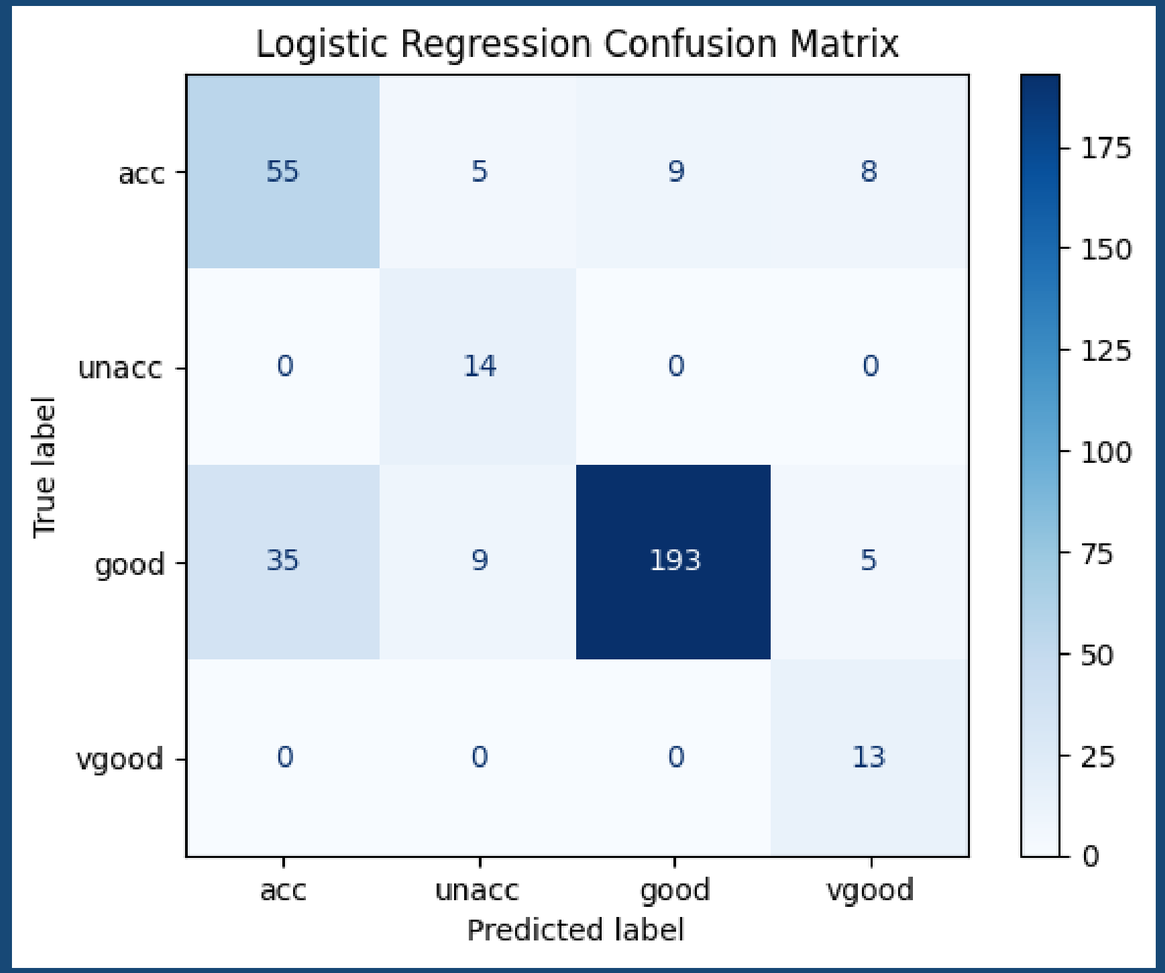
SESUDAH

```
class
acc      967
unacc    967
good     967
vgood    967
Name: count, dtype: int64
```

Perlu melakukan Oversampling dengan SMOTE agar dapat jumlah distribusi class nya dapat seimbang

LOGISTIC REGRESSION

Logistic Regression adalah metode klasifikasi untuk memprediksi probabilitas apakah sesuatu masuk ke dalam salah satu kategori(Ya/tidak atau 1/0).



Logistic Regression Accuracy: 0.7947976878612717

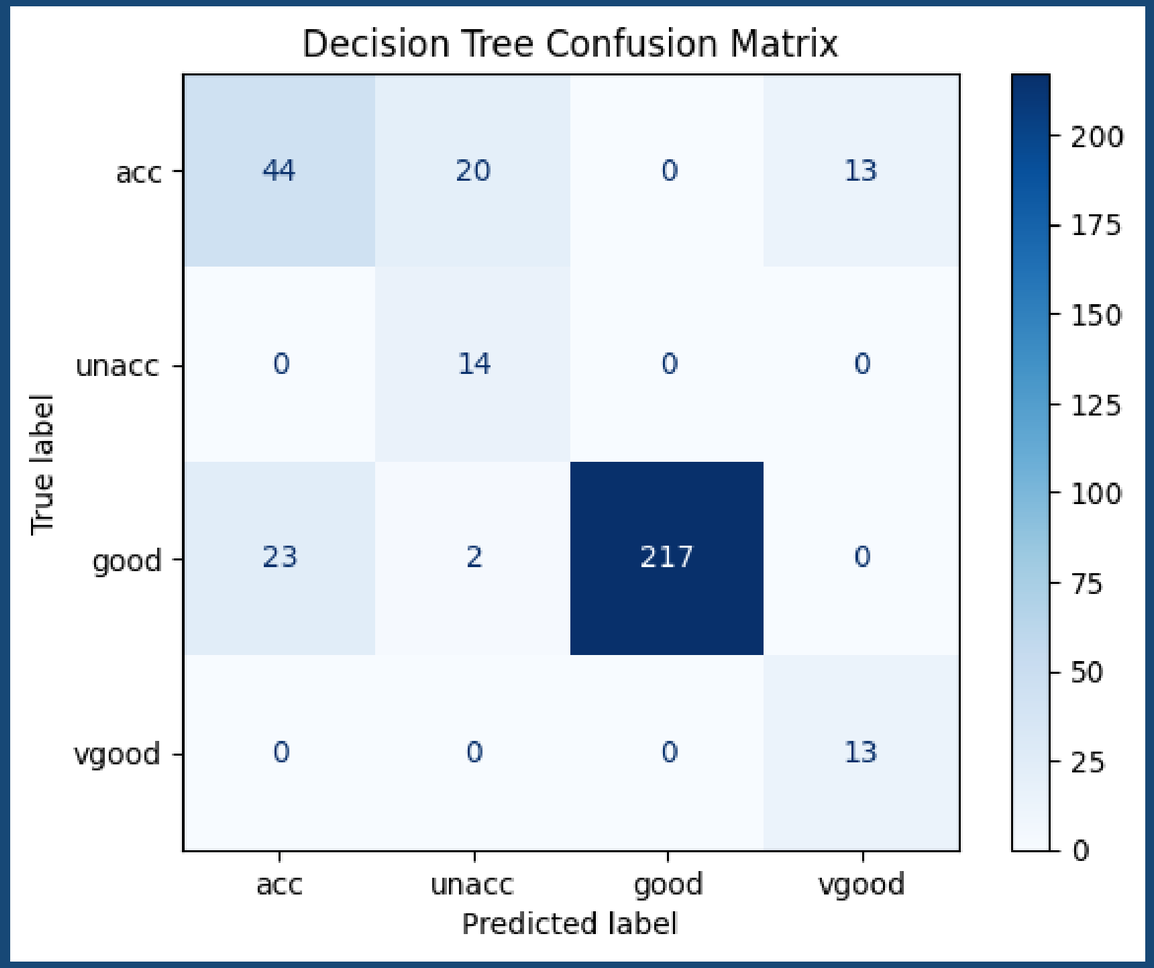
	precision	recall	f1-score	support
acc	0.61	0.71	0.66	77
good	0.50	1.00	0.67	14
unacc	0.96	0.80	0.87	242
vgood	0.50	1.00	0.67	13
accuracy			0.79	346
macro avg	0.64	0.88	0.72	346
weighted avg	0.84	0.79	0.81	346

Class acc - TP: 55, FP: 35, FN: 22, TN: 234
Class unacc - TP: 14, FP: 14, FN: 0, TN: 318
Class good - TP: 193, FP: 9, FN: 49, TN: 95
Class vgood - TP: 13, FP: 13, FN: 0, TN: 320

Memperoleh akurasi model 79% , Memiliki TP paling dominan prediksi kelas dengan benar di kelas good, namun banyak kesalahan prediksi

DECISION TREE

Decision Tree Claasfication adalah metode klasifikasi yang menggunakan struktur pohon untuk memodelkan keputusan berdasarkan fitur (input) dan target (output).

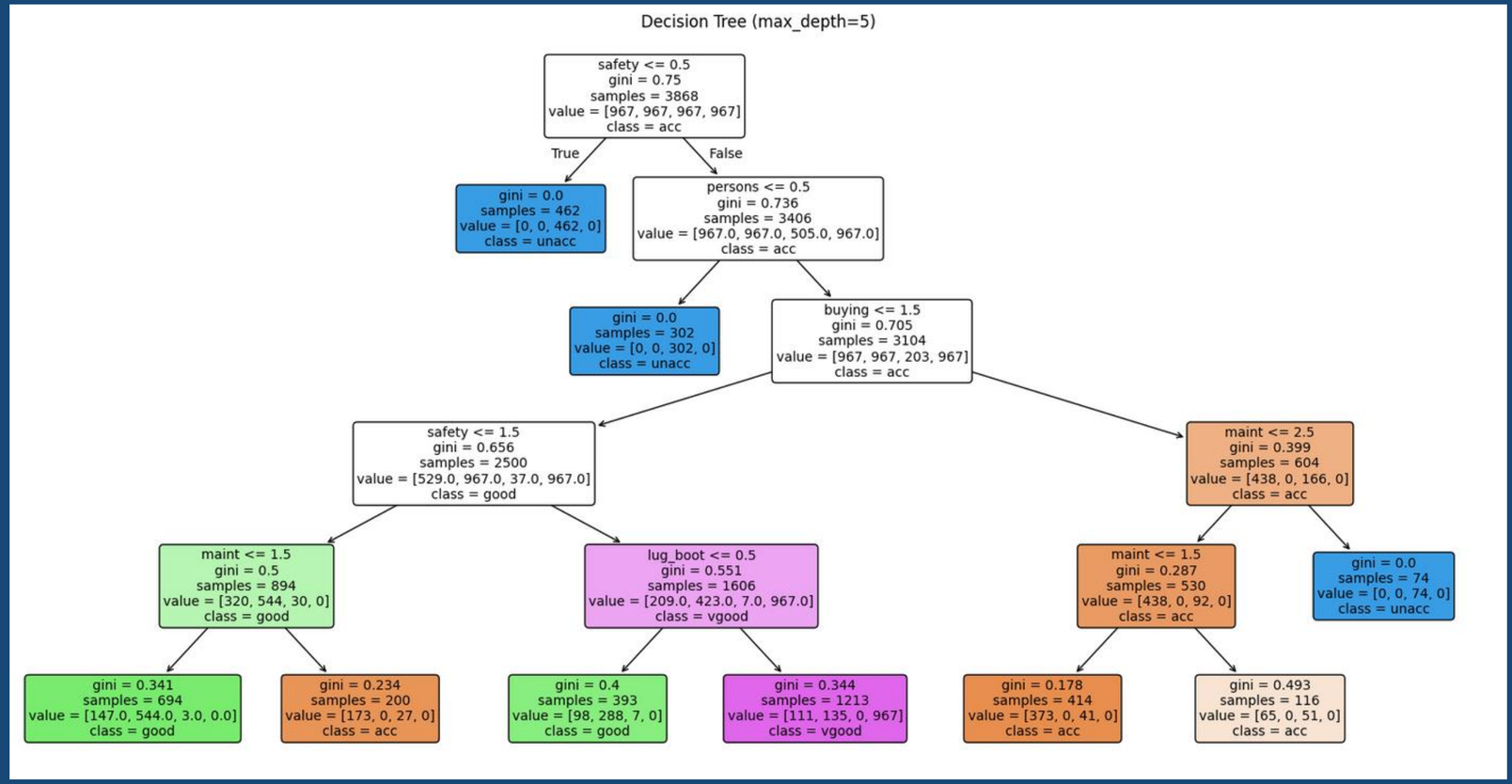


Decision Tree Accuracy: 0.8323699421965318

	precision	recall	f1-score	support
acc	0.66	0.57	0.61	77
good	0.39	1.00	0.56	14
unacc	1.00	0.90	0.95	242
vgood	0.50	1.00	0.67	13
accuracy			0.83	346
macro avg	0.64	0.87	0.70	346
weighted avg	0.88	0.83	0.85	346

Class acc - TP: 44, FP: 23, FN: 33, TN: 246
Class unacc - TP: 14, FP: 22, FN: 0, TN: 310
Class good - TP: 217, FP: 0, FN: 25, TN: 104
Class vgood - TP: 13, FP: 13, FN: 0, TN: 320

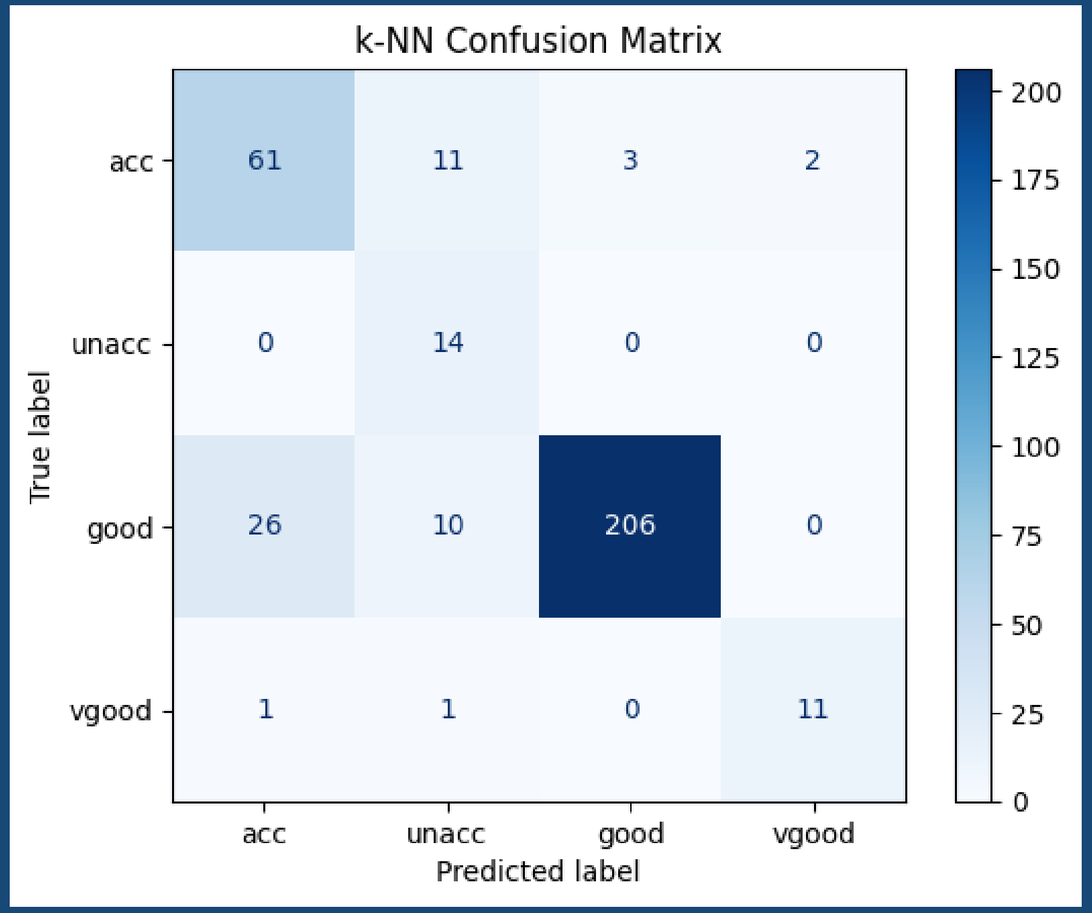
memperoleh Akurasi model 83%, paling dominan memprediksi benar di kelas good, masih banyak kesalahan prediksi kelas



fitur safety paling penting dan menjadi split pertama (akar pohon), diikuti oleh fitur persons, buying, maint, dan lug_boot. fitur-fitur tersebut sangat memengaruhi keputusan klasifikasi. Beberapa fitur nilai gini nya 0 (di node biru) yang berarti Node sangat murni (homogen), semua data berasal dari satu kelas. Namun ada beberapa yang tidak homogen gininya lebih dari 0 (selain biru) yang berarti ada data bercampur dari beberapa kelas.

K-NN

K-NN Classifcation adalah algoritma machine learning yang bekerja dengan cara membandingkan data baru dengan data yang sudah ada di dataset berdasarkan jarak terdekatnya

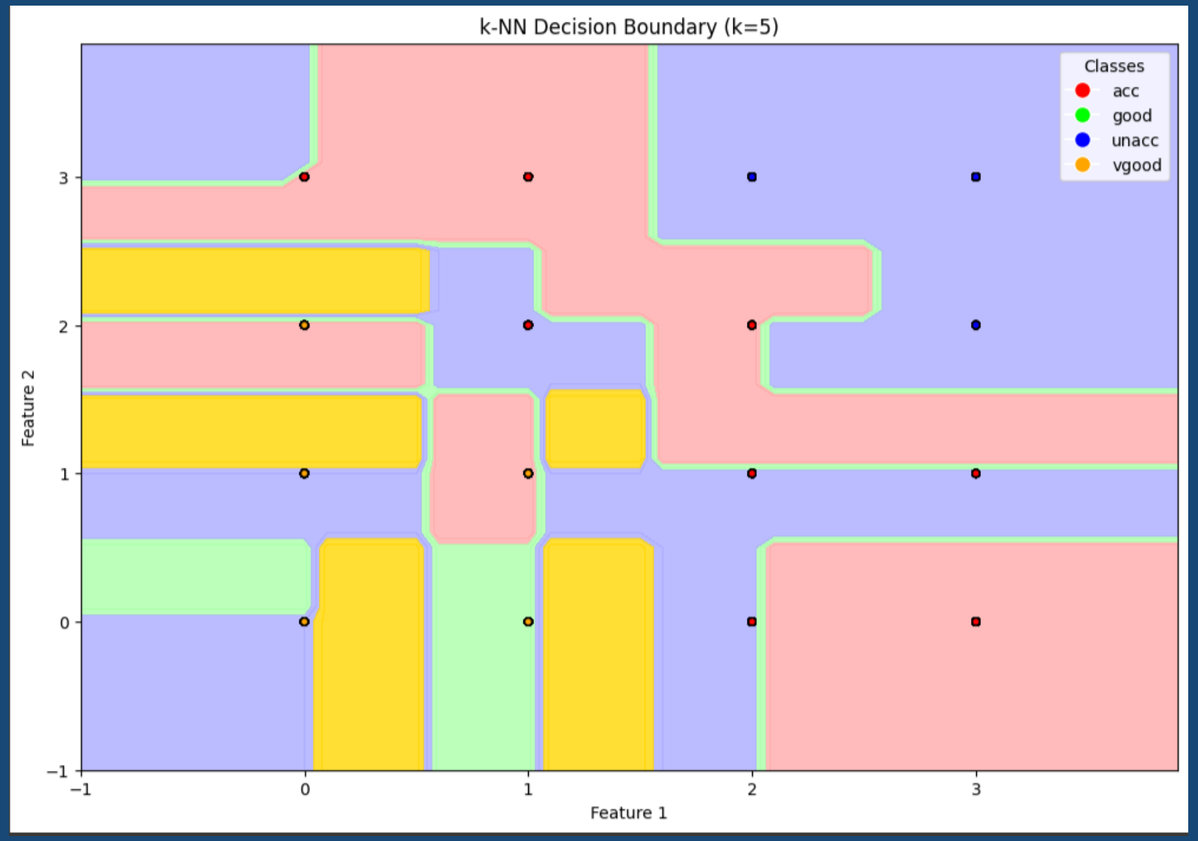


k-NN Accuracy: 0.8439306358381503

	precision	recall	f1-score	support
acc	0.69	0.79	0.74	77
good	0.39	1.00	0.56	14
unacc	0.99	0.85	0.91	242
vgood	0.85	0.85	0.85	13
accuracy			0.84	346
macro avg	0.73	0.87	0.76	346
weighted avg	0.89	0.84	0.86	346

Class acc - TP: 61, FP: 27, FN: 16, TN: 242
Class unacc - TP: 14, FP: 22, FN: 0, TN: 310
Class good - TP: 206, FP: 3, FN: 36, TN: 101
Class vgood - TP: 11, FP: 2, FN: 2, TN: 331

Memperoleh akurasi model 84% , paling dominan memprediksi benar di kelas good, masih ada kesalahan prediksi kelas

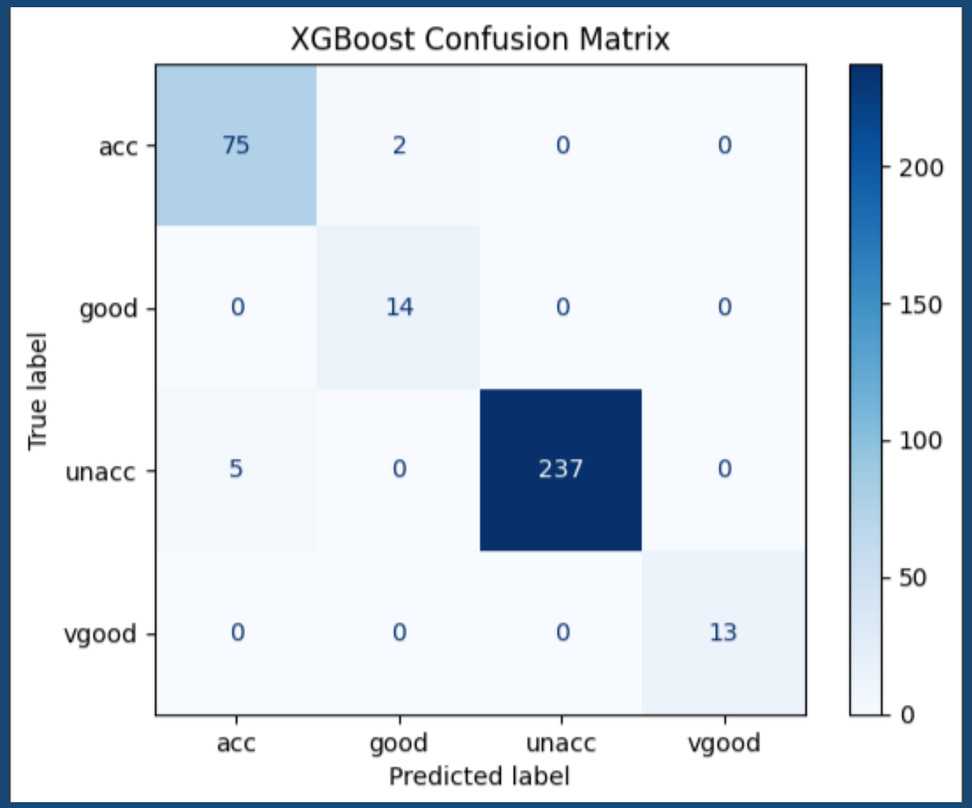


Model bekerja dengan baik, karena sebagian besar titik sesuai dengan warna area.
Namun ada Kesalahan prediksi mungkin terjadi di area perbatasan antara warna contohnya pada titik kuning(vgood) ada di area unacc

XGBOOST

CLASSIFICATION

XGBoost Classification adalah metode machine learning yang memanfaatkan algoritma gradient boosting untuk secara efisien dan akurat memprediksi kelas atau kategori dengan membangun pohon keputusan secara iteratif.

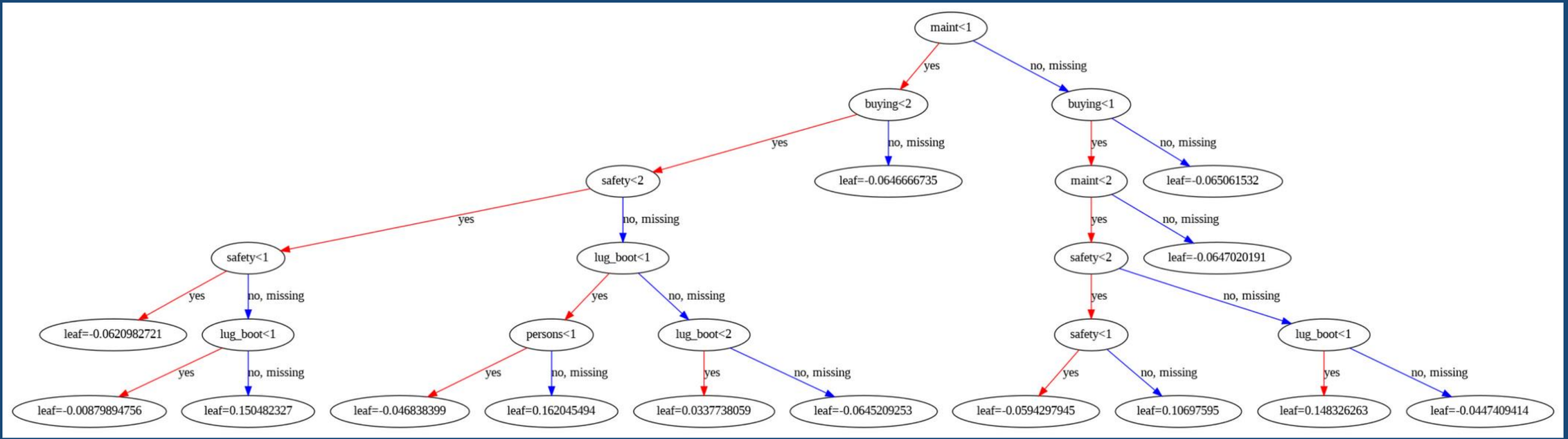


XGBoost Accuracy: 0.9797687861271677

	precision	recall	f1-score	support
acc	0.94	0.97	0.96	77
good	0.88	1.00	0.93	14
unacc	1.00	0.98	0.99	242
vgood	1.00	1.00	1.00	13
accuracy			0.98	346
macro avg	0.95	0.99	0.97	346
weighted avg	0.98	0.98	0.98	346

Class acc - TP: 75, FP: 5, FN: 2, TN: 264
Class unacc - TP: 14, FP: 2, FN: 0, TN: 330
Class good - TP: 237, FP: 0, FN: 5, TN: 104
Class vgood - TP: 13, FP: 0, FN: 0, TN: 333

Memperoleh akurasi model 97%, paling dominan prediksi benar di kelas unacc dan memiliki minimnya kesalahan prediksi diantara pipeline lainnya



buying adalah fitur yang paling berpengaruh dalam tree ini, diikuti oleh persons, lug_boot, dan doors. Model sederhana dengan jalur keputusan yang jelas, tetapi tetap efektif memisahkan data.

KESIMPULAN

Berdasarkan hasil evaluasi dapat disimpulkan bahwa untuk kasus dataset ini pipeline yang memiliki hasil atau akurasi terbaik adalah XGBOOST di akurasi 97% dan memiliki kinerja yang hampir sempurna di semua metrik yang berarti lebih cocok untuk kasus dataset ini. Kemudian untuk kasus dataset ini dapat menunjukkan model sangat mampu menangani data yang tidak seimbang serta memiliki generalisasi yang sangat baik.

Link Youtube:
https://youtu.be/eQQOHM-w3EE

THANK YOU

THANK YOU
THANK YOU