Assignment #4

Modeling & Feature Importance

Random Forest Classification

DS5G05 - Nada Salsabila - G

Problem Statement

Pendekatan yang tepat untuk dataset ini adalah menggunakan metode supervised learning, karena dalam kasus dataset heart disease terdapat data yang sudah diberi label (apakah pasien menderita penyakit jantung atau tidak). Model yang digunakan adalah klasifikasi dengan menggunakan algoritma Random Forest.

Pendekatan tersebut dirasa tepat karena melalui pembelajaran supervised dengan model klasifikasi, kita dapat memprediksi apakah seseorang menderita penyakit jantung berdasarkan data fitur yang tersedia.

Modeling

Random Forest Classifier

- Splitting Data
- Inisialisasi Model
- Perhitungan Akurasi
- Perbandingan dengan Algoritma lainnya

Random Forest Classifier

Random Forest adalah salah satu algoritma terbaik dalam machine learning. Algoritma ini menggunakan decision tree atau pohon keputusan untuk melangsungkan proses seleksi, di mana tree atau pohon decision tree akan dibagi secara rekursif berdasarkan data pada kelas yang sama. Dalam hal ini, penggunaan tree yang semakin banyak akan memengaruhi akurasi yang didapat menjadi lebih optimal.

Source:
https://algorit.ma/blog/cara-kerja-algoritma-random-forest-2022/

Splitting Data

Tahapan splitting dimulai dengan persiapan data terlebih dahulu menjadi dua kategori yaitu fitur dan target. Fitur(x) merupakan variabel independen yang digunakan oleh model untuk membuat prediksi, sedangkan target(y) adalah variabel dependen yang akan diprediksi oleh model.

```
[26]  1 X = df2.drop('target', axis=1)
      2 y = df2['target']
      3
      4 # Membagi data menjadi data latih dan data uji
      5 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Rasio yang digunakan saat melakukan split data adalah 20:80. Dengan 20% dari data digunakan sebagai data uji dan 80% lainnya digunakan sebagai data latih

Inisialisasi Model

```
# Inisialisasi model Random Forest
random_forest = RandomForestClassifier(random_state=42)

# Melatih model menggunakan data latih
random_forest.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Model yang digunakan adalah Random Forest Classifier

Tahapan ini dilakukan untuk menginisialisasi model dan melakukan pelatihan pada data latihnya.

Akurasi Model

```
Akurasi: 0.86
Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.88      0.81      0.84        26
           1       0.85      0.90      0.88        31

    accuracy                           0.86        57
   macro avg       0.86      0.86      0.86        57
weighted avg       0.86      0.86      0.86        57
```

Insight

Model memiliki akurasi yang cukup baik (86%), menunjukkan bahwa model dapat mengklasifikasikan sebagian besar data dengan benar.

Meskipun presisi dan recall kedua kelas cukup tinggi, **presisi** kelas 1 **lebih rendah dibandingkan** dengan **recall** kelas 1. Ini mungkin mengindikasikan bahwa ada beberapa instance kelas 1 yang diprediksi sebagai kelas 0 (false negative) dan model ini perlu ditingkatkan untuk mengurangi kesalahan ini.

Perbandingan Model

dengan Algoritma Naive Bayes

Alasan

Dapat dilihat bahwasannya nilai akurasi yang dihasilkan oleh algoritma naive bayes sangat kecil, berbanding terbalik dengan algoritma Random Forest.

Berdasarkan hal tersebut, pada assignment kali ini pertimbangan model yang saya gunakan adalah Random Forest, karena nilai akurasi yang dihasilkan jauh lebih besar

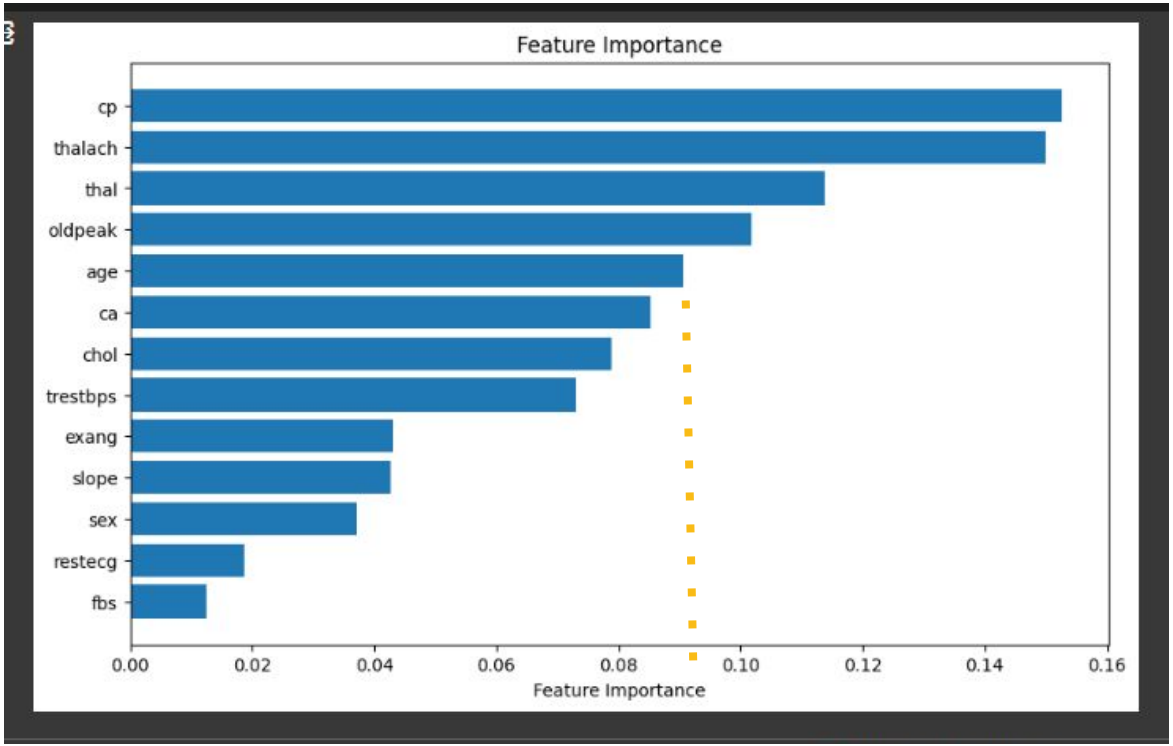
```
Akurasi model Naive Bayes: 0.40350877192982454
              precision    recall  f1-score   support

           0       0.32      0.32      0.32        25
           1       0.47      0.47      0.47        32

    accuracy                           0.40        57
   macro avg       0.39      0.39      0.39        57
weighted avg       0.40      0.40      0.40        57
```

Feature Importance

- Cek Feature Importance
- Cek Performa Model berdasarkan Feature Importance



Feature Importance

Dari plot di samping dapat diketahui terdapat **5 feature** dengan jumlah batas terbanyak.

Yaitu: feature cp, thalach, thal, oldpeak, dan age

Penjelasan Feature Importance

Terdapat 3 fitur yang paling membantu model kita untuk membedakan pasien dengan penyakit jantung atau tidak, yaitu cp, thalach, dan thal.

- cp: "**jenis nyeri jantung**" adalah salah satu fitur yang sering digunakan untuk memprediksi risiko penyakit jantung atau kondisi medis lainnya
- thalach: sering digunakan sebagai parameter penting dalam pengukuran **kebugaran jantung** dan sering diukur selama tes stres jantung untuk mengevaluasi seberapa baik jantung seseorang berfungsi selama aktivitas fisik yang intens.
- thal: adalah Thalium Stress Test yang mengukur bagaimana **darah mengalir ke jantung**

Berdasarkan penjelasan tiap fitur di atas tidak heran jika fitur tersebut sangat berpengaruh terhadap model prediksi, karena selain memang saling **memiliki korelasi**, dalam dunia medis ketiga fitur tersebut juga menjadi **indikator penentu** apakah seseorang terjangkit penyakit jantung atau tidak.

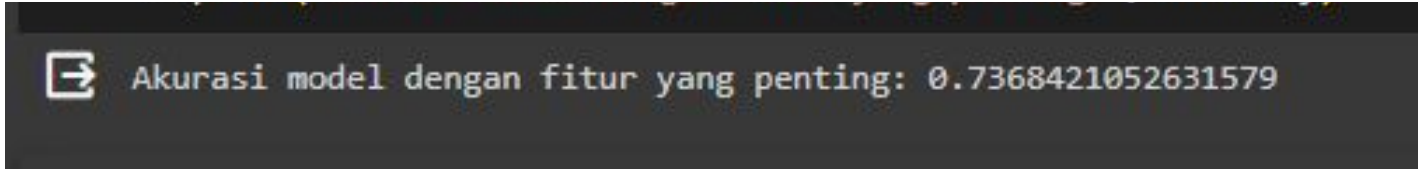
Performa Model berdasarkan Feature Importance

```
# Tentukan ambang batas untuk feature importance
threshold = 0.09  # Ganti dengan ambang batas yang sesuai
```

Untuk melakukan pengecekan akurasi model berdasarkan feature importance yang telah ditemukan, kita perlu menggunakan metode **threshold** sebagai penentu ambang batas dari feature yang dimiliki

Nilai threshold yang digunakan adalah **0.09** hal ini berarti feature importance di atas ambang batas tersebut akan diambil untuk mendapatkan indeks dari fitur-fitur yang memenuhi kondisi tersebut.

Output Akurasi Model



Insight

Ternyata, akurasi model yang dihasilkan setelah melakukan perhitungan performa berdasarkan 5 feature importance dari dataset heart disease, hanya didapatkan nilai akurasi sebesar 0,736...

Dimana nilai akurasi tersebut **jauh lebih kecil dibandingkan dengan nilai akurasi dari keseluruhan model sebelumnya.**

Output Akurasi Model

Kesimpulan

Meskipun nilai akurasi dari model dengan feature importance jauh lebih kecil, tidak dapat dikatakan bahwasannya performa model dengan hanya menggunakan beberapa subset tidak optimal.

Justru, dengan akurasi 0.736.... merupakan jumlah akurasi yang cukup optimal. Mengingat bahwasannya feature yang digunakan hanya mengambil 5 subset teratas. Hal ini berarti, **dengan hanya melalui 5 feature tersebut, sebuah model dapat memprediksi dengan baik apakah seseorang dapat diklasifikasikan memiliki penyakit jantung atau tidak.**

The End

Saya siap menunggu feedback dari kakak mentor, Terima Kasih