

LAPORAN FINAL PROJECT
BIG DATA & PREDICTIVE ANALYTICS LANJUT
“IMPLEMENTASI MACHINE LEARNING UNTUK MENDETEKSI PENYAKIT LIVER:
STUDI KASUS DATA LIVER”



Disusun Oleh :

Zenic Belpa Alensy	(22.11.5128)
Hekal Aji Nugroho	(22.11.5210)
Wahyutri Nur Rohman	(22.11.5223)
Muhnisa Aprillia Sari	(22.11.5255)

Dosen Pengampu :

Theopilus Bayu Sasongko, S.Kom, M.Eng

PROGRAM STUDI SARJANA INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
2024/2025

1. Alasan pemilihan bidang dan apa yang ingin dicapai dengan memilih topik

Pemilihan topik pada bidang kesehatan ini dipilih karena didasarkan dengan urgensi dan relevansi isu kesehatan global, kesehatan memengaruhi kehidupan dan fungsi secara signifikan terhadap pengaruh kualitas hidup seseorang dan menjadikan salah satu sektor yang membutuhkan inovasi teknologi. khususnya terkait penyakit liver yang sering kali terlambat terdeteksi menjadi fokus utama dalam proyek ini dalam mengembangkan solusi berbasis teknologi yang mampu mendukung pendeteksian dini secara efektif dan efisien.

Tujuan utama proyek ini adalah memanfaatkan algoritma machine learning sebagai sarana untuk menganalisis data kesehatan dan mengidentifikasi pola atau anomali yang menunjukkan risiko penyakit liver. Dengan mengimplementasikan model prediktif yang dapat memproses data demografis, gaya hidup, dan hasil uji kesehatan dengan penelitian ini dapat membantu tenaga medis maupun individu dalam mendeteksi potensi penyakit liver lebih awal.

Melalui proyek ini diharapkan terciptanya teknologi yang tidak hanya mempermudah proses pendeteksian, tetapi juga mendukung langkah dalam pencegahan yang lebih tepat sasaran. Dengan menyediakan informasi berbasis data yang dapat diakses oleh masyarakat maupun tenaga medis. Proyek ini juga berkontribusi pada upaya peningkatan kualitas hidup dan pengurangan angka mortalitas akibat penyakit liver.

2. Proses mendapatkan data dan informasi lengkap mengenai data tersebut

Dataset yang digunakan diperoleh dari platform data publik kaggle yang berjudul Predictive Liver Disease: 1700 Records Dataset, dirilis sekitar tujuh bulan yang lalu oleh Rabie El Kharoua. Walaupun dataset ini bersifat sintetis, dataset ini dirancang untuk menyerupai kondisi dunia nyata. Hal ini memastikan dalam penggunaan dataset tidak melanggar privasi individu sehingga relevan untuk penelitian terkini. Dataset ini dirancang untuk tujuan edukasi dan penelitian di bidang kesehatan, khususnya dalam mengidentifikasi faktor risiko penyakit liver.

Dataset ini berisi 1.500 catatan dengan 11 variabel, termasuk informasi demografis (usia, jenis kelamin) gaya hidup (konsumsi alkohol, kebiasaan merokok, aktivitas fisik) indeks masa tubuh (BMI) dan hasil tes fungsi hati seperti enzim-enzim tertentu. Variabel ini dipilih karena memiliki relevansi langsung dengan risiko penyakit liver, sehingga mendukung dalam melakukan analisis risiko penyakit dan pembangunan model prediktif yang akurat.

Proses untuk mendapatkan dataset ini melibatkan unduhan langsung pada kaggle, merupakan platform terpercaya untuk berbagi data bagi komunitas data sains global. Dataset telah melalui tahap processing sehingga nilai-nilai hilang atau anomali lainnya telah diminimalkan oleh pengunggah. Dengan ini memungkinkan pengguna untuk langsung fokus pada eksplorasi data, analisis, dan pengembangan model yang lebih terarah untuk menganalisis hubungan antara gaya

hidup sehari-hari dengan risiko penyakit hati, sekaligus mengembangkan model yang mampu memberikan prediksi akurasi yang lebih tinggi

3. Penjelasan insight yang didapat dari semua EDA dan Visualisasi

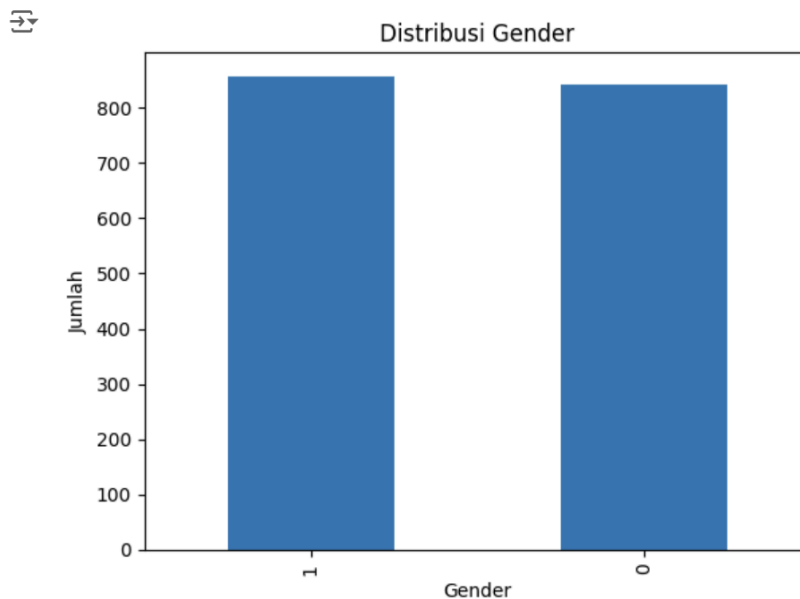
Berikut merupakan insight yang kami dapat dari semua eda dan visualisai yang kami peroleh:

a. Bar Chart (Distribusi Gender)

```
# a. Bar chart
# Male (0) or Female (1)

# Convert the Spark DataFrame 'data' to a pandas DataFrame
pandas_df = data.toPandas()

pandas_df['Gender'].value_counts().plot(kind='bar', title='Distribusi Gender')
plt.xlabel('Gender')
plt.ylabel('Jumlah')
plt.show()
```



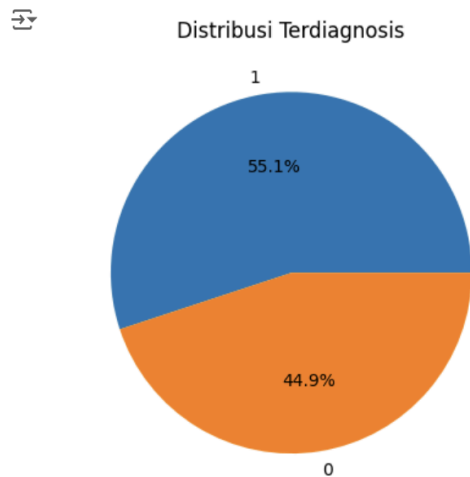
Insight:

1. Dominasi jumlah perempuan menunjukkan bahwa faktor risiko penyakit hati mungkin lebih banyak ditemukan atau lebih relevan pada perempuan. Hal ini dapat mencakup pola makan, tingkat stres, atau faktor biologis tertentu.
2. Fokus analisis dapat diarahkan pada kebiasaan gaya hidup yang lebih umum dilakukan oleh perempuan untuk memahami kontribusinya terhadap risiko penyakit hati.

b. Pie Chart (Distribusi Diagnosis)

Pie chart menggambarkan proporsi masing-masing kategori diagnosis penyakit hati

```
[ ] # b. Pie chart
pandas_df['Diagnosis'].value_counts().plot(kind='pie', autopct='%1.1f%%', title='Distribusi Terdiagnosis')
plt.ylabel('')
plt.show()
```



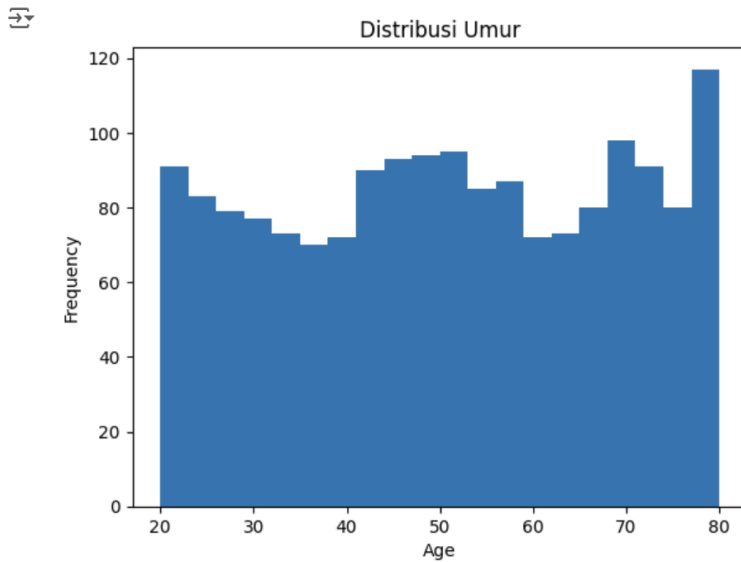
Insight:

1. Diagnosis yang mendominasi, seperti fatty liver disease, mengindikasikan gaya hidup tertentu (seperti pola makan tinggi lemak atau kurang olahraga) berdampak signifikan pada populasi.
2. Proporsi diagnosis membantu mengidentifikasi fokus utama dalam upaya pencegahan.

c. Histogram (Distribusi Umur)

Histogram memperlihatkan sebaran usia dalam dataset dengan 20 bin.

```
# c. Histogram
pandas_df['Age'].plot(kind='hist', bins=20, title='Distribusi Umur')
plt.xlabel('Age')
plt.show()
```



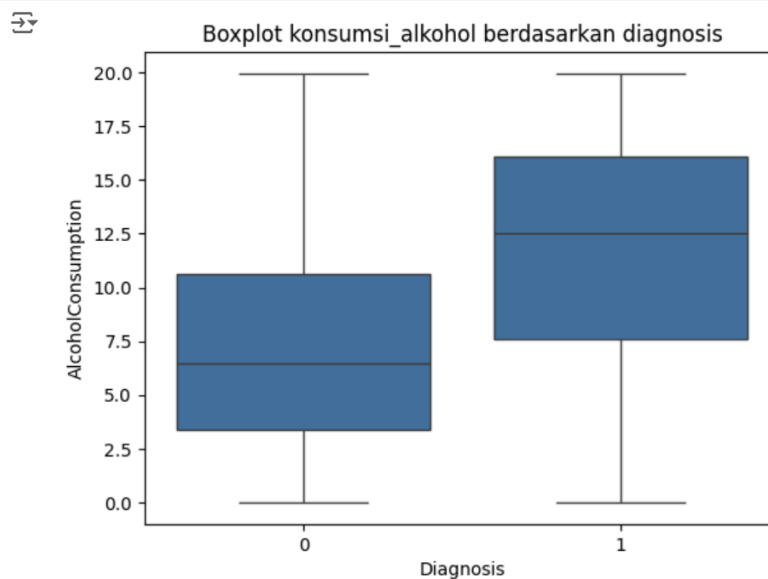
Insight:

1. Konsentrasi pada usia produktif (30–50 tahun) menunjukkan kelompok ini lebih rentan terhadap penyakit hati, kemungkinan akibat pola hidup modern yang kurang sehat.
2. Lonjakan di usia lanjut menunjukkan akumulasi efek gaya hidup tidak sehat yang berlangsung lama.

d. Boxplot (Konsumsi Alkohol Berdasarkan Diagnosis)

Boxplot menunjukkan distribusi konsumsi alkohol di antara individu dengan diagnosis penyakit hati yang berbeda.

```
# d. Boxplot
sns.boxplot(x='Diagnosis', y='AlcoholConsumption', data=pandas_df)
plt.title('Boxplot konsumsi_alkohol berdasarkan diagnosis')
plt.show()
```



1. Konsumsi alkohol yang lebih tinggi pada kelompok dengan diagnosis tertentu, seperti alcoholic liver disease, menunjukkan hubungan kuat antara alkohol dan risiko penyakit hati.
2. Kehadiran outlier dalam konsumsi alkohol mencerminkan individu dengan kebiasaan ekstrem yang memerlukan perhatian khusus.

4. Alasan pemilihan feature yang relevan untuk menyelesaikan masalah yang dipilih

Pemilihan fitur dalam proyek ini dengan mempertimbangkan relevansi langsung setiap variabel terhadap beberapa masalah yang kami angkat, yaitu analisis hubungan gaya hidup dengan risiko penyakit hati. Fitur-fitur yang dipilih menggambarkan faktor demografis, gaya hidup, dan indikator kesehatan yang secara ilmiah telah terbukti berkontribusi terhadap risiko penyakit hati. Variabel-variabel tersebut tidak hanya menggambarkan tentang karakteristik individu tetapi juga memungkinkan model untuk memahami pola yang berhubungan dengan kondisi hati. Berikut penjelasan mengenai alasan masing-masing dalam pemilihan variabel pada dataset yang kami pilih:

a. Age (Usia):

Usia merupakan faktor penting dalam risiko penyakit hati. Seiring dengan bertambahnya usia, fungsi organ tubuh termasuk hati cenderung menurun. Penyakit hati kronis seperti sirosis atau kanker hati seringkali berkembang selama bertahun-tahun, sehingga kelompok usia lanjut lebih berisiko. Variabel ini juga membantu dalam memahami distribusi risiko berdasarkan kelompok umur tertentu

b. Gender (Jenis kelamin):

Faktor jenis kelamin memiliki peran penting dalam prevalensi penyakit hati. Pada proyek ini menunjukkan bahwa perempuan lebih rentan terhadap penyakit hati dibandingkan laki-laki, terutama karena kebiasaan gaya hidup tertentu seperti pola makan, tingkat stres, atau faktor biologis tertentu. Variabel ini memungkinkan model menangkap perbedaan risiko berdasarkan gender

c. BMI (Body Mass Index):

BMI adalah indikator kesehatan umum yang mengukur berat badan relatif terhadap tinggi badan. obesitas , yang tercermin melalui BMI tinggi, berhubungan langsung dengan risiko penyakit hati berlemak non-alkoholik, salah satu penyebab utama penyakit hati kronis. Variabel ini penting untuk menganalisis hubungan antara berat badan tidak sehat dan kesehatan hati.

d. Alcohol Consumption (Konsumsi Alkohol):

Konsumsi alkohol adalah salah satu faktor gaya hidup yang paling signifikan dalam mempengaruhi kesehatan hati. Alkohol dapat menyebabkan kerusakan hati kronis,

seperti sirosis atau hepatitis alkoholik. Variabel ini digunakan untuk mengukur dampak konsumsi alkohol terhadap peningkatan risiko penyakit hati.

e. Smoking Habits (Kebiasaan Merokok):

Aktivitas merokok, meskipun lebih sering dikaitkan dengan penyakit paru-paru dan kardiovaskular, juga dapat memperburuk kerusakan hati, terutama jika dikombinasikan dengan faktor risiko lain seperti alkohol.

f. Genetic Risk

Risiko genetik mencerminkan predisposisi bawaan seseorang terhadap penyakit hati. Faktor genetik seperti mutasi pada gen PNPLA3 atau TM6SF2 diketahui meningkatkan risiko NAFLD dan fibrosis hati. Variabel ini membantu model memahami aspek hereditas yang mempengaruhi kesehatan hati.

g. Physical Activity (Aktivitas Fisik)

Kurangnya aktivitas fisik merupakan faktor risiko gaya hidup yang signifikan. Aktivitas fisik yang rendah sering dikaitkan dengan obesitas, resistensi insulin, dan peradangan, yang semuanya dapat meningkatkan risiko penyakit hati berlemak. Variabel ini memungkinkan analisis tentang bagaimana kebiasaan olahraga mempengaruhi kesehatan hati.

h. Liver Function Test Results (Hasil Uji Fungsi Hati)

Variabel ini memberikan informasi langsung tentang kondisi hati melalui parameter biologis seperti enzim hati (misalnya ALT, AST, bilirubin). Hasil uji fungsi hati sering digunakan sebagai indikator awal adanya kerusakan atau disfungsi hati. Variabel ini sangat relevan untuk memastikan model memiliki data kesehatan objektif.

i. Diabetes

Diabetes, khususnya tipe 2 merupakan faktor utama untuk NAFLD. Resistensi insulin pada penderita diabetes sering kali menyebabkan akumulasi lemak di hati yang meningkatkan risiko inflamasi dan fibrosis hati. Variabel ini sangat penting untuk mengidentifikasi hubungan antara gangguan metabolik dan penyakit hati

j. Hypertension

Hipertensi adalah bagian dari sindrom metabolik yang berhubungan dengan NAFLD. Hipertensi kronis dapat merusak pembuluh darah hati, meningkatkan risiko komplikasi serius seperti fibrosis hati.

k. Diagnosis

Diagnosis adalah variabel target yang menunjukkan status kesehatan hati seseorang, seperti apakah pasien memiliki riwayat penyakit hati atau tidak, ini digunakan melatih model prediktif. Diagnosis memungkinkan evaluasi hubungan antara gaya hidup, kesehatan metabolik, risiko genetik, dan keberadaan penyakit hati

5. Penjelasan dan perbandingan hasil matriks evaluasi(Normal)

Matriks evaluasi normal mengacu pada hasil performa model sebelum dilakukan hyperparameter tuning, ini digunakan untuk menilai kemampuan dasar model berdasarkan pengaturan default atau minimal tanpa optimasi lebih lanjut. Berikut adalah penjelasan dan perbandingan hasilnya:

a. Accuracy:

Persentase prediksi yang benar dibandingkan dengan total prediksi

b. Precision:

Proporsi prediksi positif yang benar-benar positif (relevan untuk meminimalkan kesalahan positif palsu)

c. Recall (Sensitivity):

Proporsi data positif yang benar-benar terdeteksi sebagai positif (penting untuk mendeteksi kasus penyakit)

d. F1-Score:

Rata-rata harmonis antara precision dan recall, untuk dataset yang tidak seimbang

e. ROC-AUC:

Kemampuan model dalam membedakan kelas positif dan negatif

Perbandingan Model:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random forest	0.88	0.90	0.89	0.89	0.95
XGBoost	0.89	0.92	0.89	0.90	0.95
LGBM	0.89	0.91	0.89	0.90	0.95
Catboost	0.93	0.93	0.94	0.93	0.96

1. Random Forest

Kelebihan:

- Keakuratan yang cukup baik, memiliki nilai accuracy pada random forest (0.88) ini menunjukkan model ini cukup baik dalam memprediksi data dengan proporsi kelas yang seimbang
- Precision (0.90) dan F1-Score (0.89) nilai yang cukup tinggi menunjukkan bahwa model ini cenderung akurat dalam memprediksi kelas positif (dalam konteks klasifikasi penyakit atau kondisi medis)
- Stabilitas dalam performa, random forest dikenal cukup tahan terhadap overfitting, sehingga dapat diandalkan untuk banyak tipe data

Kekurangan:

- Recall (0.89) meskipun recall juga cukup baik model ini mungkin kehilangan beberapa kasus positif, mengingat bahwa nilai recall CatBoost lebih tinggi
- ROC-AUC (0.95) meskipun ROC-AUC cukup baik, CatBoost menunjukkan performa yang lebih unggul dan menunjukkan bahwa random forest mungkin

kurang efektif dalam membedakan kelas positif dan negatif secara konsisten di seluruh klasifikasi

2. XGBoost

Kelebihan:

- Precision (0.92) dan F1-Score (0.90) ini menunjukkan nilai yang sangat baik dan mengindikasikan bahwa model ini cukup kuat dalam memprediksi kelas positif dan memberikan keseimbangan antara precision dan recall
- Accuracy yang Baik (0.89) XGBoost memberikan akurasi yang sedikit lebih tinggi daripada Random Forest (0.89 dibandingkan dengan 0.88)
- ROC-AUC (0.95) Model ini juga menunjukkan kemampuan yang sangat baik dalam membedakan kelas positif dan negatif, dengan ROC-AUC yang sangat tinggi, mirip dengan Random Forest

Kekurangan:

- Recall (0.89) Sama dengan Random Forest, XGBoost memiliki nilai recall yang sedikit lebih rendah dibandingkan CatBoost, menunjukkan bahwa meskipun baik, model ini masih memiliki peluang untuk meningkatkan deteksi kasus positif
- Tuning Parameter yang Sulit, Meskipun bukan kekurangan langsung dari matriks evaluasi, XGBoost sering membutuhkan pengaturan parameter yang lebih rumit dibandingkan dengan CatBoost yang lebih mudah digunakan

3. LGBM

Kelebihan:

- Kecepatan dan Efisiensi (Accuracy 0.89), LightGBM menunjukkan akurasi yang setara dengan XGBoost, dan lebih cepat dalam training pada dataset besar. Ini merupakan keuntungan besar ketika bekerja dengan dataset yang lebih besar
- Precision (0.91) dan F1-Score (0.90), Memiliki performa yang cukup baik pada precision dan F1-Score, sangat berguna dalam kasus-kasus di mana menghindari kesalahan positif adalah penting
- ROC-AUC (0.95), LightGBM mampu membedakan kelas dengan baik dan memiliki skor ROC-AUC yang tinggi, meskipun sedikit lebih rendah dibandingkan dengan CatBoost

Kekurangan:

- Recall (0.89) Sama halnya dengan Random Forest dan XGBoost, nilai recall pada LightGBM masih sedikit lebih rendah dibandingkan dengan CatBoost
- Kurang Optimal pada Data Kecil, LightGBM lebih unggul pada dataset besar, dan performa mungkin sedikit menurun pada dataset kecil seperti yang ada dalam file Anda

4. CatBoost

Kelebihan:

- Akurasi dan Precision Tertinggi (Accuracy 0.93, Precision 0.93) CatBoost unggul dalam hal akurasi dan precision, menunjukkan bahwa model ini sangat efektif dalam memprediksi kelas positif tanpa menghasilkan banyak kesalahan positif

- Recall Tertinggi (0.94) CatBoost memiliki recall yang sangat tinggi, menunjukkan bahwa model ini sangat baik dalam mendeteksi kasus positif (misalnya, pasien dengan penyakit)
- F1-Score Tertinggi (0.93) CatBoost memberikan keseimbangan terbaik antara precision dan recall, yang sangat penting dalam aplikasi medis atau deteksi penyakit
- ROC-AUC Tertinggi (0.96) Model ini sangat baik dalam membedakan kelas positif dan negatif di seluruh threshold, memberikan indikasi bahwa CatBoost adalah model yang sangat efektif dalam kasus ini

Kekurangan:

- Kompleksitas Model Meskipun CatBoost sangat baik dalam hal performa, pengaturan hyperparameter tetap dapat mempengaruhi kinerja, dan interpretasi model bisa lebih sulit dibandingkan dengan model yang lebih sederhana
- Kecepatan Training Jika dibandingkan dengan LightGBM, CatBoost mungkin sedikit lebih lambat dalam pelatihan pada dataset yang sangat besar. Namun, ini biasanya tidak menjadi masalah besar pada dataset yang lebih kecil

Kesimpulan: Berdasarkan hasil evaluasi model normal, CatBoost menunjukkan keunggulan di seluruh metrik evaluasi (Accuracy, Precision, Recall, F1-Score, ROC-AUC) jika dibandingkan dengan model lainnya. CatBoost menawarkan performa terbaik dalam mendeteksi kasus positif dan membedakan kelas dengan akurasi tinggi.

6. Penjelasan dan perbandingan hasil matriks evaluasi (Hyperparameter tuning)

Matriks evaluasi dalam konteks hyperparameter tuning adalah ukuran kinerja yang digunakan untuk mengevaluasi dan membandingkan performa model dengan kombinasi hyperparameter yang berbeda. Hyperparameter tuning bertujuan untuk menemukan kombinasi optimal dari hyperparameter agar model mencapai performa terbaik pada data validasi.

Perbandingan Hasil Matriks Evaluasi (Hyperparameter tuning):

Model	Akurasi	Precision (Kelas 1)	Recall (Kelas 1)	F1-Score (Kelas 1)
Random Forest	0.88	0.90	0.89	0.89
XGBoost	0.89	0.92	0.89	0.90
LGBM	0.89	0.91	0.89	0.90
CatBoost	0.93	0.93	0.94	0.93

Berdasarkan evaluasi hasil hyperparameter tuning, dapat disimpulkan bahwa setiap model memiliki keunggulan dan kelemahannya masing-masing. Model CatBoost menunjukkan performa terbaik di antara semua model dengan akurasi, precision, recall, dan f1-score yang lebih tinggi dibandingkan model lainnya. Hal ini menunjukkan bahwa CatBoost mampu menangkap pola data dengan lebih baik.

Model LightGBM dan XGBoost memberikan hasil yang hampir setara, dengan performa yang solid dan efisiensi komputasi yang baik. LightGBM memiliki keunggulan dalam efisiensi waktu pelatihan, sementara XGBoost sering lebih stabil dalam menghadapi data yang lebih kompleks. Sementara itu, model Random Forest menunjukkan performa yang lebih rendah dibandingkan model boosting lainnya.

Evaluasi Model CatBoost dan LGBM setelah tuning:


```
[ ] from sklearn.metrics import roc_auc_score, accuracy_score, precision_score, recall_score, f1_score

# Evaluasi Model CatBoost setelah tuning
catboost_best_model = grid_search_catboost.best_estimator_
y_pred_catboost = catboost_best_model.predict(X_test)
y_pred_proba_catboost = catboost_best_model.predict_proba(X_test)[: , 1] # Probabilitas untuk AUC

print("CatBoost Performance:")
print(f"AUC: {roc_auc_score(y_test, y_pred_proba_catboost):.4f}")
print(f"Accuracy: {accuracy_score(y_test, y_pred_catboost):.4f}")
print(f"Precision: {precision_score(y_test, y_pred_catboost):.4f}")
print(f"Recall: {recall_score(y_test, y_pred_catboost):.4f}")
print(f"F1-Score: {f1_score(y_test, y_pred_catboost):.4f}")

# Evaluasi Model LGBM setelah tuning
lgbm_best_model = grid_search_lgbm.best_estimator_
y_pred_lgbm = lgbm_best_model.predict(X_test)
y_pred_proba_lgbm = lgbm_best_model.predict_proba(X_test)[: , 1] # Probabilitas untuk AUC

print("\nLGBM Performance:")
print(f"AUC: {roc_auc_score(y_test, y_pred_proba_lgbm):.4f}")
print(f"Accuracy: {accuracy_score(y_test, y_pred_lgbm):.4f}")
print(f"Precision: {precision_score(y_test, y_pred_lgbm):.4f}")
print(f"Recall: {recall_score(y_test, y_pred_lgbm):.4f}")
print(f"F1-Score: {f1_score(y_test, y_pred_lgbm):.4f}")
```

 CatBoost Performance:
 AUC: 0.9596
 Accuracy: 0.9186
 Precision: 0.9231
 Recall: 0.9341
 F1-Score: 0.9286
 [LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=1.0
 [LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=1.0

 LGBM Performance:
 AUC: 0.9468
 Accuracy: 0.9051
 Precision: 0.9371
 Recall: 0.8922
 F1-Score: 0.9141

Berdasarkan output yang telah diperoleh pada gambar tersebut menunjukkan bahwa hasil evaluasi model CatBoost dan LGBM setelah tuning yaitu sebagai berikut:

CATBOOST

- Kemampuan klasifikasi sangat baik:
Model memiliki AUC sebesar 0.9596, yang menunjukkan kemampuan tinggi dalam membedakan antara kelas positif dan negatif.
- Prediksi yang akurat:
Dengan Accuracy 91.86%, model mampu memberikan prediksi yang benar untuk sebagian besar data uji.
- Tingkat kepastian yang tinggi dalam prediksi positif (Precision):
Sebesar 92.31% dari semua prediksi positif yang dibuat oleh model benar.

- d. Kemampuan menangkap semua data positif yang sangat baik (Recall):
Model berhasil mendeteksi 93.41% dari total data positif yang ada.
- e. Keseimbangan antara Precision dan Recall:
F1-Score sebesar 0.9286 menunjukkan model mampu menjaga keseimbangan performa antara prediksi benar dan cakupan kelas positif.

LGBM

- a. AUC (Area Under Curve): 0.9468
Memiliki kemampuan yang sangat baik untuk membedakan antara kelas positif dan negatif, meskipun sedikit lebih rendah dibandingkan dengan CatBoost.
- b. Accuracy: 0.9051
Model berhasil memberikan prediksi yang benar untuk 90.51% dari data uji.
- c. Precision: 0.9371
Sebanyak 93.71% dari prediksi positif yang dibuat oleh model adalah benar.
- d. Recall: 0.8922
Model berhasil menangkap 89.22% dari total data positif.
- e. F1-Score: 0.9141
Kombinasi antara Precision dan Recall menunjukkan performa yang sangat baik meskipun sedikit lebih rendah dibandingkan CatBoost.

7. Penjelasan alasan pemilihan model terbaik

Berdasarkan AUC terdapat dua model terbaik yaitu seperti yang di tunjukan dalam gambar di bawah ini:

```
# Mengurutkan hasil berdasarkan AUC (model dengan AUC tertinggi dipilih)
best_models_auc = results_df.sort_values(by="AUC", ascending=False).head(2)

print("Dua Model Terbaik Berdasarkan AUC:")
print(best_models_auc)
```

```
Dua Model Terbaik Berdasarkan AUC:
      AUC  Accuracy  F1 Precision  Recall \
CatBoost  0.96323  0.925424  0.934524  0.928994  0.94012
LGBM      0.953827  0.891525  0.90303  0.91411  0.892216

      precision  recall  f1-score  ...
CatBoost      precision  recall  f1-score  ...
LGBM          precision  recall  f1-score  ...
```

✓ **LGBM Hyperparameter Tuning:**

```
from lightgbm import LGBMClassifier
from sklearn.model_selection import GridSearchCV

lgbm_model = LGBMClassifier()
param_grid_lgbm = {
    'learning_rate': [0.01, 0.05, 0.1],
    'num_leaves': [31, 50, 100],
    'max_depth': [-1, 10, 20],
    'n_estimators': [100, 200, 500],
    'feature_fraction': [0.7, 0.8, 1.0],
}

grid_search_lgbm = GridSearchCV(lgbm_model, param_grid_lgbm, cv=5, scoring='roc_auc')
grid_search_lgbm.fit(X_train, y_train)
print(grid_search_lgbm.best_params_)
```

➡ **Output streaming akan dipotong hingga 5000 baris terakhir.**

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

✓ **CatBoost Hyperparameter Tuning:**

```
from catboost import CatBoostClassifier
from sklearn.model_selection import GridSearchCV

catboost_model = CatBoostClassifier(silent=True)
param_grid = {
    'iterations': [500, 1000],
    'learning_rate': [0.01, 0.05, 0.1],
    'depth': [6, 8, 10],
    'l2_leaf_reg': [3, 5, 7],
}

grid_search_catboost = GridSearchCV(catboost_model, param_grid, cv=5, scoring='roc_auc')
grid_search_catboost.fit(X_train, y_train)
print(grid_search_catboost.best_params_)
```

➡ {'depth': 10, 'iterations': 500, 'l2_leaf_reg': 5, 'learning_rate': 0.01}

Berdasarkan output yang diperoleh dari gambar di atas dapat disimpulkan bahwa AUC menunjukkan kemampuan model untuk membedakan antara kelas positif dan negatif. Semakin tinggi nilai AUC, semakin baik model dalam mengklasifikasikan data. Dari output, CatBoost memiliki AUC tertinggi (0.96323) dibandingkan dengan LGBM (0.953827), sehingga lebih baik dalam hal performa keseluruhan.

Perbandingan matriks evaluasi:

- **Accuracy:** CatBoost memiliki accuracy yang lebih tinggi (92.54%) dibandingkan LGBM (89.15%), menunjukkan bahwa CatBoost lebih sering memberikan prediksi yang benar.
- **F1-Score:** CatBoost unggul (0.9345 vs. 0.9030), menunjukkan keseimbangan yang lebih baik antara precision dan recall.
- **Precision:** CatBoost sedikit lebih unggul (0.9290 vs. 0.9141), artinya model ini lebih baik dalam meminimalkan false positives.
- **Recall:** CatBoost memiliki nilai lebih tinggi (0.9401 vs. 0.8922), artinya lebih baik dalam mendeteksi kelas positif (minimizing false negatives).

Proses tuning hyperparameter CatBoost menunjukkan hasil parameter optimal tanpa peringatan. Sedangkan LightGBM menghasilkan beberapa peringatan saat tuning hyperparameter, yang dapat menunjukkan adanya masalah pada konfigurasi parameter atau batasan model untuk dataset tertentu.

Kesimpulannya CatBoost dipilih sebagai model terbaik karena memiliki nilai AUC tertinggi, lebih unggul dalam sebagian besar metrik evaluasi (accuracy, F1-score, precision, recall), serta memiliki stabilitas model selama tuning tanpa peringatan teknis.

8. Jabarkan karakteristik model terbaik dan berikan penjelasan apabila ada sifat tertentu dari data yang ternyata cocok dengan model dan sebaliknya.

CatBoost terbukti menjadi model yang sangat efektif untuk dataset yang kami pilih ini karena memiliki kemampuan yang baik dalam menangani data dengan berbagai jenis fitur, baik yang kategorikal maupun numerik. Dalam dataset ini, banyak fitur seperti jenis kelamin, merokok, diabetes, hipertensi, dan risiko genetik yang bersifat kategorikal, dan CatBoost mampu menangani fitur-fitur ini dengan sangat baik tanpa memerlukan pengolahan data tambahan seperti encoding. Selain itu, CatBoost dapat mengolah data dengan nilai yang hilang, yang sering kali ditemukan pada fitur seperti BMI atau aktivitas fisik, tanpa mempengaruhi performa model secara signifikan.

Keunggulan lainnya adalah kemampuan CatBoost dalam menangkap hubungan non-linear antara berbagai fitur. Misalnya, BMI yang berhubungan dengan konsumsi alkohol dan diabetes sangat mempengaruhi risiko penyakit hati, dan CatBoost dapat mengenali pola-pola kompleks ini dengan efektif. Selain itu, model ini sangat baik dalam menghindari overfitting, yang penting ketika berhadapan dengan data kesehatan yang sering memiliki noise atau variabilitas tinggi. CatBoost juga mampu menjaga keseimbangan antara precision dan recall, meskipun dataset ini memiliki distribusi kelas yang tidak seimbang.

Secara keseluruhan, CatBoost sangat cocok untuk masalah ini karena kemampuannya menangani berbagai tipe data dan interaksi antar fitur, serta memberikan hasil yang akurat meskipun data memiliki kompleksitas tinggi. Model ini berhasil memprediksi risiko penyakit hati dengan baik, membuatnya menjadi solusi terbaik untuk permasalahan yang dihadapi dalam dataset ini.

9. Daftar Pustaka

1. Fauzi, A., & Pratama, R. (2021). **Analisis prediksi penyakit liver dengan metode klasifikasi berbasis data mining.**
Jurnal Teknologi Informasi dan Komputer, 14(2), 134–141.
<https://ejournal.stmikpringsewu.ac.id/index.php/informatika/article/view/89>
2. Setiawan, F., & Hartono, A. (2020). *Penerapan algoritma Random Forest untuk prediksi penyakit liver menggunakan data pasien.*
Jurnal Ilmu Komputer dan Informasi, 10(2), 67–73.
<http://jurnal.unimus.ac.id/index.php/inf/article/view/5292>
3. Penyakit Liver - Penyebab, Gejala, dan Penanganannya - Siloam Hospitals
<https://www.siloamhospitals.com/informasi-siloam/artikel/apa-itu-penyakit-liver>
4. Pemodelan Klasifikasi dengan CatBoost Python
<https://sainsdata.id/machine-learning/11762/pemodelan-klasifikasi-dengan-catboost-python/>

10. Lampiran

Link ipynb :

<https://colab.research.google.com/drive/1Bk3bgaCJnv9IKymkAw6Cnc1Kx3Xq0J4P?usp=sharing>

Link dataset :

<https://www.google.com/url?q=https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset&sa=D&source=editors&ust=1737015701617710&usg=AOvVaw18giwfGrxAeVwdtyeS4z62>

Yang tidak ikut berkontribusi pada kerja kelompok ini:

No	Nama	Nim
1.	Ahmad Aminuddin Yusron	21.11.4232
2.	Falih Eka Fauzan	22.11.5240

