IMPLEMENTASI XGBOOST DAN RANDOM FOREST PADA PENYAKIT STUNTING

Ria Wulandari
202010370311168

Data Science
Universitas Muhammadiyah Malang

ABSTRAK

Stunting merupakan masalah serius di Indonesia, mempengaruhi sekitar 21,6% balita akibat kekurangan gizi kronis. Penelitian ini fokus pada implementasi metode XGBoost dan Random Forest untuk mengidentifikasi faktor risiko stunting. Penelitian ini penting untuk pengembangan strategi pencegahan yang lebih efektif terhadap stunting, mempertimbangkan faktor gizi, ekonomi, dan lingkungan.

LATAR BELAKANG

Stunting adalah masalah kesehatan global yang mendesak, terutama di negara berkembang seperti Indonesia. Data Kementerian Kesehatan 2022 menunjukkan 21,6% balita Indonesia mengalami stunting, yang disebabkan oleh kurangnya asupan gizi kronis pada anak usia 0-23 bulan. Faktor-faktor seperti nutrisi ibu hamil, kondisi ekonomi, dan lingkungan turut memengaruhi kondisi ini. Dampak stunting tidak hanya mempengaruhi masa anak-anak, tetapi juga memiliki efek jangka panjang, termasuk menurunkan kecerdasan (IQ) dan meningkatkan risiko penyakit kronis di masa dewasa. Penelitian ini bertujuan mengeksplorasi penggunaan algoritma XGBoost dan Random Forest untuk meningkatkan akurasi dan kehandalan klasifikasi penderita stunting. Algoritma ini dikenal efektif dalam menangani hubungan nonlinier, missing values, dan overfitting, yang menjadikannya alat yang unggul dalam analisis data kesehatan. Implementasi XGBoost dan Random Forest diharapkan memberikan kontribusi signifikan dalam memahami faktor-faktor yang mempengaruhi stunting, serta menghasilkan model prediktif yang lebih baik dibandingkan dengan metode sebelumnya, seperti SVM dan KNN

RUMUSAN MASALAH

1. Bagaimana implementasi metode XGBoost dan Random Forest dalam klasifikasi penderita stunting pada anak dapat meningkatkan akurasi dan kehandalan model Berbeda dengan metode Support Vector Machine (SVM) dan KNN ?
2. Apa saja faktor-faktor atau atribut yang paling berpengaruh terhadap klasifikasi penderita stunting menggunakan metode tersebut ?
3. Berapa nilai akurasi yang di dapatkan dengan menggunakan metode XGBoost dan Random Forest ?

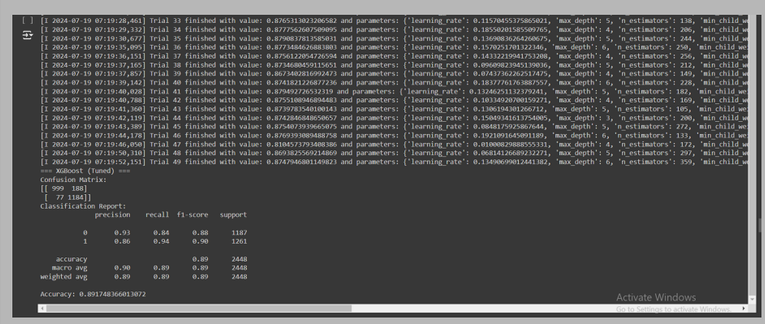
TUJUAN MASALAH

Untuk mengevaluasi hasil kerja klasifikasi metode XGBoost dan Random Forest dengan berupa accuracy dan memberikan model prediksi untuk klasifikasi penderita penyakit stunting.

METHODOLOGY

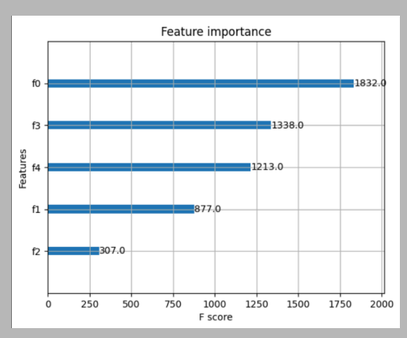
Untuk mengevaluasi hasil kerja klasifikasi metode XGBoost dan Random Forest dengan berupa accuracy dan memberikan model prediksi untuk klasifikasi penderita penyakit stunting.

BEST HYPERPARAMETER



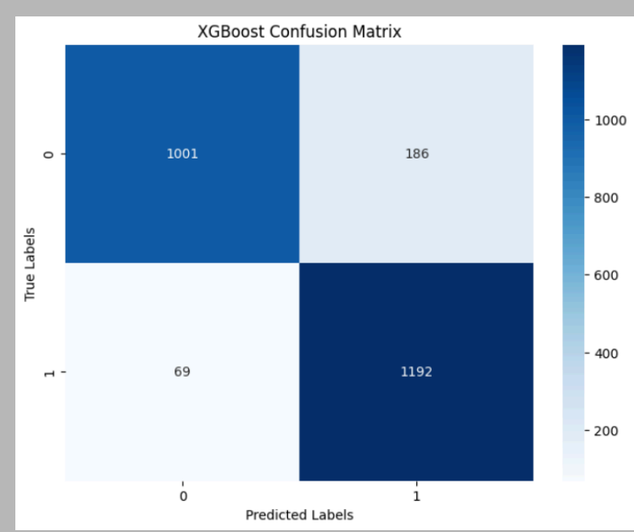
```
[ ] [I 2024-07-19 07:19:28,461] Trial 33 finished with value: 0.876531303206582 and parameters: {'learning_rate': 0.115704553758650 21, 'max_depth': 5, 'n_estimators': 138, 'min_child_w
[I 2024-07-19 07:19:29,332] Trial 34 finished with value: 0.877756260750909 5 and parameters: {'learning_rate': 0.185502015855097 65, 'max_depth': 4, 'n_estimators': 206, 'min_child_w
[I 2024-07-19 07:19:30,677] Trial 35 finished with value: 0.879083783150831 and parameters: {'learning_rate': 0.169008362642060 75, 'max_depth': 5, 'n_estimators': 244, 'min_child_w
[I 2024-07-19 07:19:35,095] Trial 36 finished with value: 0.877348462688380 3 and parameters: {'learning_rate': 0.157025170132234 6, 'max_depth': 6, 'n_estimators': 258, 'min_child_we
[I 2024-07-19 07:19:36,151] Trial 37 finished with value: 0.875612205472659 4 and parameters: {'learning_rate': 0.143322195417532 08, 'max_depth': 4, 'n_estimators': 256, 'min_child_w
[I 2024-07-19 07:19:37,165] Trial 38 finished with value: 0.873468045015565 1 and parameters: {'learning_rate': 0.090498239451390 36, 'max_depth': 5, 'n_estimators': 212, 'min_child_w
[I 2024-07-19 07:19:37,857] Trial 39 finished with value: 0.867340281699247 3 and parameters: {'learning_rate': 0.074373622625174 75, 'max_depth': 4, 'n_estimators': 149, 'min_child_w
[I 2024-07-19 07:19:39,142] Trial 40 finished with value: 0.874182122687236 and parameters: {'learning_rate': 0.181777617638875 57, 'max_depth': 6, 'n_estimators': 228, 'min_child_w
[I 2024-07-19 07:19:40,028] Trial 41 finished with value: 0.874892726532319 and parameters: {'learning_rate': 0.132462511237924 1, 'max_depth': 5, 'n_estimators': 182, 'min_child_we
[I 2024-07-19 07:19:40,788] Trial 42 finished with value: 0.875518094680448 3 and parameters: {'learning_rate': 0.103349207001592 71, 'max_depth': 4, 'n_estimators': 169, 'min_child_we
[I 2024-07-19 07:19:41,360] Trial 43 finished with value: 0.879978354010014 3 and parameters: {'learning_rate': 0.130619430126671 2, 'max_depth': 5, 'n_estimators': 105, 'min_child_we
[I 2024-07-19 07:19:42,119] Trial 44 finished with value: 0.874284068058657 and parameters: {'learning_rate': 0.154049416137540 75, 'max_depth': 3, 'n_estimators': 200, 'min_child_we
[I 2024-07-19 07:19:43,389] Trial 45 finished with value: 0.875407393966507 5 and parameters: {'learning_rate': 0.084817592586764 4, 'max_depth': 5, 'n_estimators': 272, 'min_child_we
[I 2024-07-19 07:19:44,170] Trial 46 finished with value: 0.876939308948758 and parameters: {'learning_rate': 0.192109164509310 9, 'max_depth': 6, 'n_estimators': 133, 'min_child_we
[I 2024-07-19 07:19:46,050] Trial 47 finished with value: 0.810457379340838 6 and parameters: {'learning_rate': 0.010008298855533 1, 'max_depth': 4, 'n_estimators': 172, 'min_child_we
[I 2024-07-19 07:19:50,110] Trial 48 finished with value: 0.869382556921486 9 and parameters: {'learning_rate': 0.068141266892322 71, 'max_depth': 5, 'n_estimators': 297, 'min_child_we
[I 2024-07-19 07:19:52,151] Trial 49 finished with value: 0.874796801149023 and parameters: {'learning_rate': 0.134906090124413 82, 'max_depth': 6, 'n_estimators': 359, 'min_child_we
=== XGBoost (Tuned) ===
Confusion Matrix:
[[ 999  188]
 [  77 1184]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.84      0.88      1187
           1       0.86      0.94      0.90      1261

    accuracy                           0.89      2448
   macro avg       0.90      0.89      0.89      2448
weighted avg       0.89      0.89      0.89      2448

Accuracy: 0.891748366013072
```

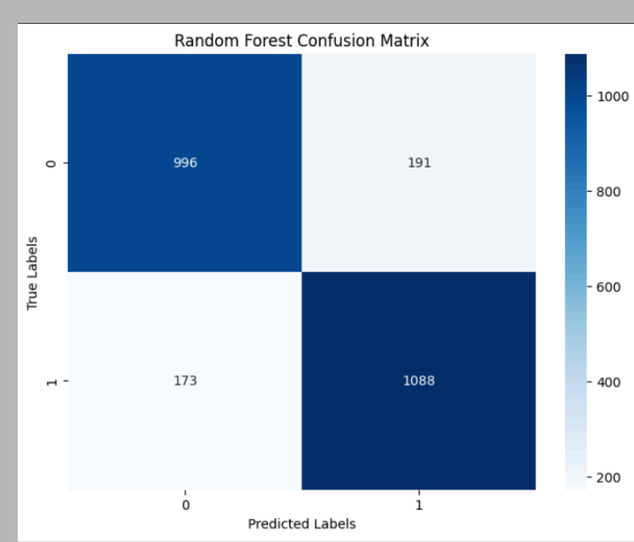
FEATURE IMPORTANCES




```
Feature: Gender, Importance: 0.20054705440998077
Feature: Age, Importance: 0.16477282345294952
Feature: Birth Weight, Importance: 0.33703359961509705
Feature: Birth Length, Importance: 0.144262433052063
Feature: Body Weight, Importance: 0.15338410437107086
Feature: Body Length, Importance: 0.0
```

CONFUSION MATRIX XGBOOST DAN RANDOM FOREST





HASIL

Untuk meningkatkan performa model XGBoost, dilakukan hyperparameter tuning menggunakan Optuna dan GridSearchCV. Optuna, dengan Bayesian optimization dan TPE, secara efisien menemukan kombinasi terbaik dari parameter yang diuji. Dalam empat kali percobaan, parameter optimal yang diperoleh meningkatkan akurasi model secara signifikan. GridSearchCV memastikan evaluasi robust melalui cross-validation pada tiga bagian data, dengan hasil tuning menunjukkan peningkatan performa model yang lebih optimal dibandingkan parameter default.

KESIMPULAN

Pada penelitian ini, XGBoost menunjukkan performa lebih baik dibandingkan Random Forest dengan akurasi 90% dibandingkan 85%. Keunggulan XGBoost terletak pada kemampuannya menangani data tidak seimbang, efisiensi pelatihan, dan regularisasi untuk mencegah overfitting. Random Forest unggul dalam interpretabilitas dan memberikan wawasan tentang pentingnya fitur, meskipun membutuhkan waktu pelatihan lebih lama. Birth Weight memiliki nilai penting tertinggi (0,33) menunjukkan pengaruh besar terhadap model. Kedua model ini efektif dalam mendukung identifikasi dini dan manajemen stunting, serta pengembangan strategi intervensi yang lebih tepat.