**PREDIKSI KESEHATAN LANSIA BERDASARKAN DATA POSYANDU DENGAN MODEL RANDOM FOREST**



**Disusun Oleh:**

**Nama   : Fauzi Hidayat**

**NIM   : A11.2022.14365**

**Kelompok : A11.4504**

**Program Studi : Sistem Informatika - S1**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS DIAN NUSWANTORO**

**SEMARANG 2024**

**Ringkasan:**

Proyek ini bertujuan untuk memprediksi kondisi kesehatan lansia berdasarkan data yang diperoleh dari kegiatan posyandu. Data tersebut mencakup atribut seperti umur, berat badan, tinggi badan, lingkar perut, dan tensi darah. Model yang digunakan untuk memprediksi kondisi kesehatan lansia adalah **Random Forest Classifier**.

**Permasalahan Project:**

Data kesehatan lansia yang didapat masih cukup sedikit karena hanya lansia pada satu dusun. Permasalahan utama adalah mengklasifikasikan kondisi kesehatan lansia menjadi dua kategori: **Normal** atau **Tinggi** berdasarkan fitur-fitur yang tersedia.

**Tujuan:**

Tujuan dari proyek ini adalah:

1. Membangun model prediksi untuk kondisi kesehatan lansia (Normal atau Tinggi) berdasarkan data yang diberikan.
2. Mengidentifikasi langkah-langkah penting dalam pemrosesan data untuk memastikan kualitas data yang baik sebelum digunakan dalam model.
3. Mengukur performa model menggunakan metrik evaluasi yang relevan seperti akurasi, precision, recall, F1-score, serta visualisasi confusion matrix dan ROC Curve.

**Model atau Alur Penyelesaian:**

Proyek ini mengikuti alur penyelesaian yang terdiri dari beberapa langkah:

1. **Pengumpulan dan Pra-pemrosesan Data**: Data dikumpulkan dengan meminta data hasil pemerriksaan lansia pada posko posyandu. Membaca data dari file CSV, membersihkan data (menghapus duplikasi, menangani nilai yang hilang), menangani outlier, dan melakukan normalisasi.
2. **Pembagian Data**: Membagi dataset menjadi data pelatihan dan data pengujian.
3. **Modeling**: Melatih model Random Forest Classifier menggunakan data pelatihan.
4. **Evaluasi Model**: Mengevaluasi performa model menggunakan berbagai metrik seperti akurasi, precision, recall, F1-score, serta AUC dan ROC Curve.
5. **Visualisasi**: Menyediakan visualisasi seperti confusion matrix dan ROC curve untuk mempermudah pemahaman performa model.

**Bagan Alur Penyelesaian:**

**[Pengumpulan Data] -> [Pra-pemrosesan Data] -> [Pembagian Data] -> [Pelatihan Model] -> [Evaluasi Model] -> [Visualisasi dan Analisis Hasil]**

3. **Penjelasan Dataset, EDA, dan Proses Fitur Dataset**

#### ****Penjelasan Dataset****:

Dataset yang digunakan berasal dari data posyandu lansia yang berisi informasi mengenai individu lansia, termasuk:

* **No** : Nomer urut (tidak digunakan dalam model).
* **Nama**: Nama individu (tidak digunakan dalam model).
* **Umur**: Umur individu lansia.
* **Berat Badan**: Berat badan individu lansia.
* **Tinggi Badan**: Tinggi badan individu lansia.
* **Lingkar Perut**: Lingkar perut individu lansia.
* **Tensi**: Kondisi kesehatan individu, berupa **Normal** atau **Tinggi** (label target).

#### ****Exploratory Data Analysis (EDA)****:

Sebelum membangun model, kami melakukan analisis awal terhadap dataset untuk memahami distribusi dan kualitas data:

* **Missing Values**: Kami mengidentifikasi kolom dengan nilai yang hilang dan mengisi nilai yang hilang menggunakan metode imputation (rata-rata untuk data numerik dan mode untuk data kategorikal).
* **Outliers**: Menggunakan metode Interquartile Range (IQR) untuk mendeteksi dan menghapus outlier pada fitur-fitur numerik.
* **Distribusi Data**: Dilakukan normalisasi pada fitur-fitur numerik seperti umur, berat badan, tinggi badan, lingkar perut, dan tensi darah untuk mempersiapkan data agar sesuai dengan model.

#### ****Proses Fitur Dataset****:

* **Pembersihan Data**: Menghapus kolom no dan nama karena tidak relevan untuk analisis.
* **Imputation**: Mengisi nilai yang hilang pada kolom numerik dengan rata-rata dan kolom kategorikal dengan mode.
* **Penghapusan Duplikasi**: Menghapus baris yang duplikat.
* **Outlier Removal**: Menggunakan IQR untuk menghapus nilai yang tidak wajar (outliers) pada fitur numerik.
* **Normalisasi**: Fitur numerik dinormalisasi menggunakan **StandardScaler** untuk memastikan bahwa model tidak terpengaruh oleh perbedaan skala antar fitur.

4. **Proses Learning/Modeling**

#### ****Pembagian Data****:

Dataset dibagi menjadi dua bagian:

* **Training Set** (80% dari data): Digunakan untuk melatih model.
* **Test Set** (20% dari data): Digunakan untuk menguji performa model.

#### ****Modeling****:

Kami menggunakan **Random Forest Classifier**, yang merupakan ensemble method yang kuat dan cocok untuk masalah klasifikasi. Model ini dilatih pada data training dan digunakan untuk memprediksi kondisi kesehatan lansia berdasarkan fitur-fitur yang tersedia.

#### ****Langkah-langkah Pemodelan****:

1. Melatih model menggunakan data training.
2. Menggunakan model untuk memprediksi data test.
3. Menghitung akurasi model dan evaluasi lebih lanjut dengan metrik lain seperti precision, recall, F1-score.

5. **Performa Model**

Berikut adalah hasil evaluasi model:

* **Akurasi**: 0.93 (93%)
* **Precision**: 0.92
* **Recall**: 0.94
* **F1-Score**: 0.93
* **Confusion Matrix**:
  + True Positives (TP):10
  + False Positives (FP): 8
  + False Negatives (FN): 2
  + True Negatives (TN): 3

**6.** **Diskusi Hasil dan Kesimpulan**

#### ****Diskusi Hasil****

Berdasarkan hasil evaluasi model **Random Forest** yang telah diterapkan pada dataset kesehatan lansia, kita dapat mengamati beberapa hal penting terkait performa model:

1. **Akurasi Model**:
   * Model menunjukkan **akurasi 93%**, yang merupakan hasil yang cukup baik untuk kasus klasifikasi kesehatan yang kompleks seperti ini. Model ini dapat dengan efektif memisahkan antara kategori **Normal** dan **Tinggi** berdasarkan fitur-fitur seperti umur, berat badan, tinggi badan, dan lingkar perut.
   * Meskipun akurasi sangat baik, ada kemungkinan bahwa model ini bisa ditingkatkan lebih jauh, terutama dalam hal deteksi kondisi **Tinggi** yang sedikit lebih sulit diprediksi dengan baik.
2. **Precision, Recall, dan F1-Score**:
   * **Precision** dan **Recall** untuk kedua kelas **Normal** dan **Tinggi** cukup seimbang, dengan precision di atas 90% untuk kedua kategori. Ini menunjukkan bahwa model tidak hanya mampu mengklasifikasikan dengan benar, tetapi juga menghindari banyak false positives (kesalahan dalam memprediksi **Tinggi** sebagai **Normal**) dan false negatives (kesalahan dalam memprediksi **Normal** sebagai **Tinggi**).
   * Nilai **F1-Score** yang tinggi untuk kedua kategori juga menunjukkan bahwa model mempertimbangkan keseimbangan antara precision dan recall dengan baik, sehingga hasil prediksi yang dihasilkan relatif dapat diandalkan dalam aplikasi dunia nyata.
   * Namun, recall untuk kategori **Tinggi** sedikit lebih rendah dibandingkan dengan kategori **Normal**. Ini berarti bahwa model lebih mudah mengidentifikasi pasien dengan tensi **Normal** daripada pasien dengan tensi **Tinggi**, yang bisa menjadi area yang perlu diperhatikan untuk perbaikan lebih lanjut.
3. **Confusion Matrix**:
   * Confusion matrix menunjukkan bahwa model melakukan beberapa kesalahan dalam mengklasifikasikan **Tinggi** (3 false negatives), yang mengindikasikan bahwa ada beberapa individu dengan tensi tinggi yang salah diklasifikasikan sebagai **Normal**.
   * Meskipun demikian, model mampu menangani kategori **Normal** dengan sangat baik, dengan hanya 2 kesalahan false positive. Ini menunjukkan bahwa model sangat berhati-hati dalam memprediksi kondisi **Normal**.
4. **ROC Curve dan AUC**:
   * Dengan **AUC 0.94**, model menunjukkan bahwa ia memiliki kemampuan yang sangat baik untuk membedakan antara kondisi **Normal** dan **Tinggi**. AUC yang mendekati 1 menunjukkan bahwa model tidak hanya bekerja dengan baik dalam klasifikasi, tetapi juga dapat diandalkan untuk aplikasi di dunia nyata, terutama untuk deteksi otomatis pada posyandu atau fasilitas kesehatan.
   * **ROC Curve** yang tajam menunjukkan bahwa model tidak hanya mampu memisahkan dua kategori dengan baik, tetapi juga berpotensi untuk dikalibrasi lebih lanjut agar lebih sensitif dalam mendeteksi **Tinggi**.

#### ****Kesimpulan****

Berdasarkan hasil yang diperoleh dari evaluasi model, berikut adalah beberapa kesimpulan yang dapat diambil:

1. **Model Sudah Cukup Efektif**:
   * Model **Random Forest** yang dikembangkan mampu mengklasifikasikan kondisi tensi darah lansia menjadi dua kategori—**Normal** dan **Tinggi**—dengan tingkat akurasi yang tinggi (93%) dan performa yang baik dalam hal precision, recall, dan F1-Score. Model ini sudah cukup efektif untuk digunakan dalam aplikasi prediksi kondisi kesehatan di posyandu atau sistem monitoring kesehatan lainnya.
2. **Perbaikan Model**:
   * Meskipun akurasi sudah tinggi, terdapat ruang untuk perbaikan dalam mendeteksi lebih akurat kondisi **Tinggi**. Penurunan **recall** pada kelas **Tinggi** bisa diatasi dengan:
     + Menggunakan teknik **oversampling** atau **undersampling** pada dataset untuk menyeimbangkan jumlah data antara dua kelas.
     + Menambahkan fitur-fitur baru yang lebih relevan, seperti faktor gaya hidup atau riwayat medis, untuk membantu model lebih baik membedakan antara kedua kategori tersebut.
     + Menyempurnakan parameter model, seperti menggunakan grid search untuk mencari nilai hyperparameter yang optimal.