LAPORAN PROYEK: ANALISIS DAN PREDIKSI DATA PENDUDUK
Proyek Remastered 2025

Daftar Isi

---

1. Latar Belakang

- **Masalah Utama**:

    1. Dataset awal hanya berisi 20 baris, tidak mencukupi untuk analisis mendalam atau pelatihan model ML.

    2. Banyak nilai kosong (NaN) yang dapat mengganggu hasil analisis.

- **Kebutuhan**:

1. Membuat dataset yang lebih besar dan realistis agar model prediksi lebih akurat.

- **Solusi**:

    1. **Augmentasi Data** – Menambah jumlah baris menjadi 200 menggunakan sintesis berdasarkan distribusi asli.

    2. **Imputasi Nilai Kosong** – Mengisi NaN dengan teknik median (numerik) dan model klasifikasi (kategorik).

    3. **Exploratory Data Analysis (EDA)** – Visualisasi pola data sebelum modeling.

    4. **Prediksi Machine Learning** – Random Forest untuk prediksi status nikah dan pendapatan.

    5. **Web App Interaktif** – Streamlit sebagai front-end untuk prediksi individual dan massal.

---

2. Tujuan

1. **Meningkatkan Kualitas Dataset**
    - Augmentasi & imputasi sehingga dataset siap pakai.
2. **Eksplorasi Data**
    - Menyajikan insight lewat visualisasi.
3. **Membangun Model Prediksi**
    - Status nikah & pendapatan menggunakan ML.
4. **Menyediakan Aplikasi Web**
    - Antarmuka mudah digunakan dengan Streamlit.

---

3. Studi Literatur / Tinjauan Pustaka

- **Augmentasi Data**
    - Teknik seperti SMOTE umum untuk menambah sampel; di sini digunakan sintesis sederhana berdasarkan distribusi asli.
- **Imputasi Nilai Kosong**
    - *Median Imputation* untuk numerik (tahan terhadap outlier).
    - *Model-Based Imputation* (Random Forest) untuk kategorik.
- **Random Forest**
    - Classifier: memprediksi kategori (status nikah).
    - Regressor: memprediksi nilai kontinu (pendapatan).
- **Streamlit**
    - Framework Python untuk web app “zero-boilerplate”.

---

4. Metodologi

4.1 Data

- **Dataset Awal**: 20 baris, kolom `ID`, `Umur`, `Pendidikan`, `Pendapatan`, `Status Nikah`.
- **Augmentasi**
    - Umur: 17–80 tahun
    - Pendidikan: SMA, D3, S1, S2
    - Pendapatan sesuai jenjang (SMA – Rp6 jt; D3 – Rp7.5 jt; S1 – Rp9 jt; S2 – Rp12 jt)
    - Hasil: ≥200 baris

- **Imputasi**
    - Median untuk `Umur` & `Pendapatan`
    - Random Forest Classifier untuk `Status Nikah`

4.2 Algoritma

1. **RandomForestClassifier**
    - Target: `Status Nikah` (Kawin/Belum)
    - Kelebihan: tahan noise, menangani variabel kategorik.
2. **RandomForestRegressor**
    - Target: `Pendapatan`
    - Kelebihan: memodelkan hubungan non-linear.

4.3 Rancangan Program

```
penduduk_analisis_app/
├── data/
│   ├── dataset_penduduk.csv
│   ├── generated_data.csv
│   └── cleaned_data.csv
├── model/
│   ├── label_encoders.pkl
│   ├── model_status_nikah.pkl
│   └── model_pendapatan.pkl
├── output/
│   └── eda_visualization.png
├── generate_data.py
├── clean_data.py
├── eda.py
├── train_model.py
├── app.py
```

```
└── requirements.txt
```

---

5. Implementasi

5.1 Input & Output Program

- **Input**:
  - `dataset_penduduk.csv` (20 baris)
  - Kolom: `Umur`, `Pendidikan`, `Pendapatan`, `Status Nikah`
- **Output**:
  - `cleaned_data.csv` (setelah imputasi)
  - `eda_visualization.png`
  - Model `.pkl` (classifier & regressor)
  - Aplikasi Streamlit

5.2 Tampilan Program (GUI)

- **Sidebar**: Menu “Beranda”, “Statistik & Visualisasi”, “Prediksi Individual”, “Prediksi Massal”.
- **Prediksi Individual**: Form untuk `Umur` & `Pendidikan`, tampilkan hasil.
- **Prediksi Massal**: Upload CSV → model → download hasil.

5.3 Cara Penggunaan

Jalankan cleaning & training

```
python generate_data.py
python clean_data.py
python train_model.py
```

Luncurkan web app

```
streamlit run app.py
```

Prediksi Individual: Isi form → klik “Predict”.

Prediksi Massal: Unggah file → unduh `hasil_prediksi_massal.csv`.

---

6. Pengujian

- **Akurasi Status Nikah**: ~90%
- **R² Pendapatan**: ~0.85
- Uji antarmuka: pastikan tidak ada error saat upload/download.

---

7. Analisis

- **EDA Insights**:
    - Mayoritas penduduk berpendidikan SMA.
    - Pendapatan meningkat seiring jenjang pendidikan.
    - Proporsi “Kawin” lebih tinggi di jenjang S1–S2.
- **Performansi Model**:
    - Hasil prediksi sesuai pola data.

---

8. Kesimpulan & Saran

Kesimpulan

- Dataset berhasil diperluas & dibersihkan.
- Model Random Forest memadai untuk prediksi.
- Aplikasi Streamlit user-friendly.

Saran Pengembangan

1. Tambahkan fitur demografi lain (jenis kelamin, pekerjaan).
2. Integrasi database (PostgreSQL/MongoDB).
3. Sediakan REST API (FastAPI).
4. Deployment di Cloud & versi mobile.

Laporan ini disusun oleh Affif (Nourivex) · Email: affif@nourivex.tech · LinkedIn: linkedin.com/in/nourivex

---

9. Daftar Pustaka

- [ ] Scikit-learn contributors. (n.d.). *Scikit-learn: Machine learning in Python* . https://scikit-learn.org
- [ ] Streamlit contributors. (n.d.). *Streamlit: Create beautiful data apps in hours, not weeks* . https://docs.streamlit.io
- [ ] Pandas development team. (2024). *pandas: Powerful data structures for data analysis, time series, and statistics* . Version 2.2.3. https://pandas.pydata.org/docs
- [ ] Matplotlib contributors. (2024). *Matplotlib: Visualization with Python* . Version 3.9.0. https://matplotlib.org
- [ ] Seaborn contributors. (n.d.). *seaborn: Statistical data visualization* . https://seaborn.pydata.org

---

Laporan ini disusun oleh Affif (Nourivex) · Email: affif@nourivex.tech · LinkedIn: linkedin.com/in/nourivex