

# **LAPORAN PROGRES PROYEK DATA SCIENCE & DATA ANALYST**

**Judul : “Pengaruh Tingkat Pendidikan terhadap Jenis Pekerjaan Penduduk Desa  
Pangombusan”**



**Oleh :**

<b>11423001</b>	<b>Samuel Leonardo Nainggolan</b>
<b>11423004</b>	<b>Marselino Tambunan</b>
<b>11423018</b>	<b>Eduward Gilbert Simanjuntak</b>
<b>11423042</b>	<b>Franky Hamonangan Sirait</b>
<b>11423044</b>	<b>Nathanael T. J. Tampubolon</b>
<b>11423045</b>	<b>Whisnu Abraham Luciano Saragih</b>

**SARJANA TERAPAN TEKNOLOGI REKAYASA**

**PERANGKAT LUNAK (TRPL)2023**

**FAKULTAS VOKASI**

## DAFTAR ISI

LAPORAN PROGRES PROYEK.....	1
DATA SCIENCE & DATA ANALYST .....	1
BAB I PENDAHULUAN .....	3
1.1. Latar Belakang .....	3
1.2. Rumusan Masalah .....	3
1.3. Tujuan Penelitian.....	3
1.4. Manfaat Penelitian .....	4
BAB II TINJAUAN PUSTAKA .....	5
2.1 Pendidikan .....	5
2.2 Pekerjaan.....	5
2.3 Data Science.....	5
2.4 Machine Learning .....	5
2.5 Algoritma yang Digunakan .....	5
BAB III METODOLOGI PENELITIAN .....	6
3.1 Sumber Data.....	6
3.2 Tahapan Penelitian.....	6
3.3 Tools yang Digunakan.....	6
3.4 Preprocessing Data.....	6
3.5 Pembagian Data .....	7
BAB IV HASIL DAN PEMBAHASAN .....	8
4.1 Distribusi Tingkat Pendidikan.....	8
4.2 Distribusi Jenis Pekerjaan.....	8
4.3 Visualisasi Pendidikan terhadap Pekerjaan.....	8
4.4 Hasil Logistic Regression .....	8
4.5 Hasil Random Forest.....	9
4.6 Confusion Matrix.....	10
BAB V KESIMPULAN DAN SARAN.....	11
5.1 Kesimpulan.....	11
5.2 Saran.....	11
BAB IV DAFTAR PUSTAKA.....	12
LAMPIRAN.....	12

# **BAB I**

## **PENDAHULUAN**

### **1.1. Latar Belakang**

Tingkat pendidikan merupakan salah satu faktor penting yang memengaruhi kualitas sumber daya manusia. Pendidikan berperan besar dalam menentukan peluang kerja, jenis pekerjaan, serta tingkat kesejahteraan seseorang. Di Desa Pangombusan, terdapat keberagaman latar belakang pendidikan dan jenis pekerjaan penduduk yang menarik untuk dikaji secara ilmiah.

Perkembangan ilmu data (Data Science) memungkinkan analisis hubungan antara tingkat pendidikan dan pekerjaan dilakukan secara lebih objektif menggunakan teknik statistik, visualisasi, serta algoritma Machine Learning. Oleh karena itu, dalam penelitian ini dilakukan analisis pengaruh tingkat pendidikan terhadap jenis pekerjaan penduduk Desa Pangombusan dengan pendekatan Data Analytics dan Machine Learning.

### **1.2. Rumusan Masalah**

Rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana distribusi tingkat pendidikan penduduk Desa Pangombusan?
2. Bagaimana distribusi jenis pekerjaan penduduk Desa Pangombusan?
3. Apakah terdapat pengaruh antara tingkat pendidikan dan jenis pekerjaan?
4. Seberapa baik model Machine Learning dalam memprediksi jenis pekerjaan berdasarkan tingkat pendidikan?

### **1.3. Tujuan Penelitian**

Tujuan dari penelitian ini adalah:

1. Mengetahui distribusi tingkat pendidikan penduduk Desa Pangombusan.
2. Mengetahui distribusi jenis pekerjaan penduduk.
3. Menganalisis pengaruh tingkat pendidikan terhadap jenis pekerjaan.
4. Membangun model Machine Learning untuk memprediksi pekerjaan berdasarkan tingkat pendidikan.

#### **1.4. Manfaat Penelitian**

1. Memberikan gambaran hubungan antara pendidikan dan pekerjaan.
2. Menjadi referensi bagi pemerintah desa dalam perencanaan sumber daya manusia.
3. Menerapkan ilmu Data Science secara nyata pada data kependudukan.

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 Pendidikan**

Pendidikan merupakan proses pembelajaran untuk meningkatkan pengetahuan, keterampilan, dan sikap individu agar mampu berperan dalam kehidupan bermasyarakat.

#### **2.2 Pekerjaan**

Pekerjaan adalah aktivitas utama yang dilakukan seseorang untuk memperoleh penghasilan atau memenuhi kebutuhan hidup.

#### **2.3 Data Science**

Data Science adalah suatu bidang ilmu yang menggabungkan statistik, pemrograman, dan analisis data untuk memperoleh wawasan dari data.

#### **2.4 Machine Learning**

Machine Learning adalah cabang kecerdasan buatan yang memungkinkan komputer belajar dari data untuk membuat prediksi.

#### **2.5 Algoritma yang Digunakan**

1. Logistic Regression
2. Random Forest Classifier

## **BAB III**

### **METODOLOGI PENELITIAN**

Metode penelitian menggunakan Python, Jupyter Notebook, dan algoritma Machine Learning (Logistic Regression dan Random Forest).

#### **3.1 Sumber Data**

Data yang digunakan merupakan data kependudukan Desa Pangombusan yang berisi informasi:

1. Tingkat Pendidikan
2. Jenis Pekerjaan

#### **3.2 Tahapan Penelitian**

1. Pengumpulan Data
2. Data Cleaning
3. Encoding Data
4. Exploratory Data Analysis (EDA)
5. Pemodelan Machine Learning
6. Evaluasi Model

#### **3.3 Tools yang Digunakan**

1. Python
2. Jupyter Notebook / Google Colab
3. Pandas, NumPy
4. Matplotlib, Seaborn
5. Scikit-learn

#### **3.4 Preprocessing Data**

Tahapan preprocessing yang dilakukan meliputi:

1. Menghapus data kosong (missing value)
2. Encoding tingkat pendidikan menjadi data numerik

3. Menghapus kategori Pelajar karena belum termasuk tenaga kerja aktif

### 3.5 Pembagian Data

Data dibagi menjadi:

- 80% data latih (training)
- 20% data uji (testing)

```
# =====  
# 6. SPLIT DATA  
# =====  
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.2, random_state=42, stratify=y  
)
```

## **BAB IV**

### **HASIL DAN PEMBAHASAN**

#### **4.1 Distribusi Tingkat Pendidikan**

Hasil analisis menunjukkan bahwa tingkat pendidikan penduduk Desa Pangombusan didominasi oleh:

1. Tamat SLTA/Sederajat
2. Sedang SD/Sederajat
3. Sedang SLTP/Sederajat
4. Sementara jumlah penduduk dengan pendidikan S2 dan S3 relatif sangat sedikit.

#### **4.2 Distribusi Jenis Pekerjaan**

Jenis pekerjaan penduduk didominasi oleh:

1. Karyawan Perusahaan Swasta
2. Wiraswasta
3. Petani
4. Ibu Rumah Tangga
5. Kelompok Pelajar dikeluarkan dari analisis karena belum termasuk tenaga kerja aktif.

#### **4.3 Visualisasi Pendidikan terhadap Pekerjaan**

Berdasarkan grafik boxplot:

1. Petani dan Ibu Rumah Tangga didominasi oleh pendidikan dasar.
2. Karyawan Swasta didominasi oleh pendidikan menengah hingga tinggi.
3. Pegawai Negeri Sipil dan Guru Swasta mayoritas berpendidikan S1.

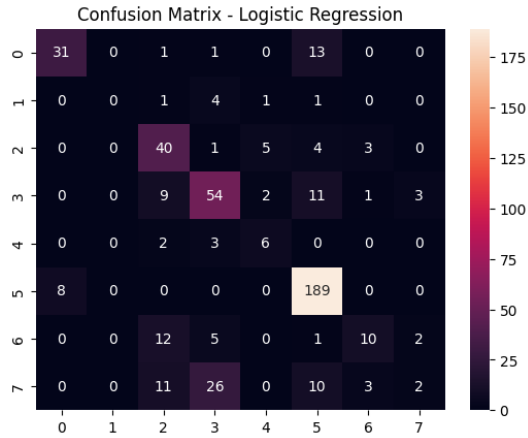
Hal ini menunjukkan bahwa semakin tinggi tingkat pendidikan, maka semakin besar peluang seseorang memperoleh pekerjaan formal dengan jenjang karier lebih baik.

#### **4.4 Hasil Logistic Regression**

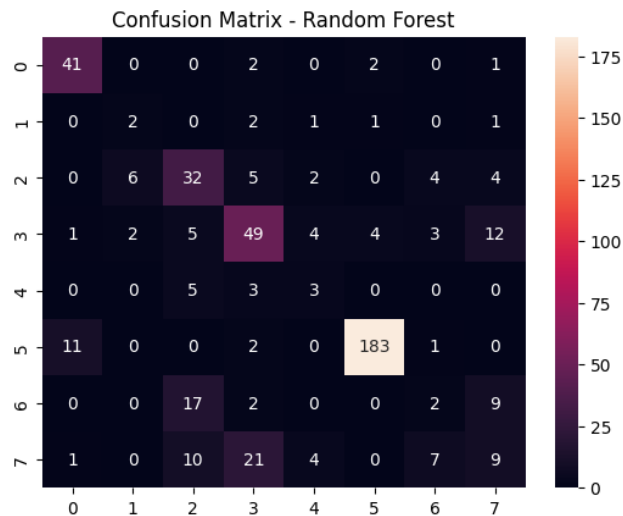
Logistic Regression adalah algoritma klasifikasi yang umum digunakan dalam machine learning, khususnya untuk memprediksi variabel target dengan tipe kategori. [1]



Model Logistic Regression digunakan untuk memprediksi jenis pekerjaan berdasarkan tingkat pendidikan dengan hasil akurasi yang cukup baik. Model mampu menangkap pola umum hubungan pendidikan terhadap pekerjaan.



#### 4.5 Hasil Random Forest



Random Forest adalah model berbasis ensemble yang menggunakan beberapa pohon keputusan untuk meningkatkan akurasi dan stabilitas, sehingga lebih efektif dalam menangani data dengan pola kompleks.[2]

Model Random Forest memberikan hasil akurasi yang lebih tinggi dibandingkan Logistic Regression. Hal ini menunjukkan bahwa hubungan antara pendidikan dan pekerjaan memiliki pola non-linear yang lebih cocok dimodelkan menggunakan Random Forest.

#### **4.6 Confusion Matrix**

Confusion matrix menunjukkan bahwa sebagian besar prediksi model berada pada kelas yang benar, meskipun masih terjadi kesalahan klasifikasi pada beberapa jenis pekerjaan yang memiliki karakteristik pendidikan yang mirip.

## **BAB V**

### **KESIMPULAN DAN SARAN**

#### **5.1 Kesimpulan**

1. Tingkat pendidikan penduduk Desa Pangombusan didominasi oleh lulusan SLTA dan pendidikan dasar.
2. Jenis pekerjaan penduduk paling banyak adalah karyawan swasta, wiraswasta, petani, dan ibu rumah tangga.
3. Terdapat pengaruh yang jelas antara tingkat pendidikan terhadap jenis pekerjaan.
4. Model Random Forest memberikan kinerja terbaik dalam memprediksi pekerjaan berdasarkan pendidikan.

#### **5.2 Saran**

1. Pemerintah desa dapat meningkatkan program pendidikan dan pelatihan kerja.
2. Masyarakat diharapkan dapat meningkatkan pendidikan untuk memperoleh pekerjaan yang lebih baik.
3. Penelitian selanjutnya dapat menambahkan variabel lain seperti usia dan pendapatan.

#### **BAB IV**

#### **DAFTAR PUSTAKA**

- [1] D. Klasifikasi and P. Kulit, “Perbandingan metode logistic regression dan random forest dalam klasifikasi penyakit kulit multikelas 1) 1,2,3),” vol. 10, no. 2, pp. 1369–1379, 2025.
- [2] A. Mu *et al.*, “Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis Hyperparameter Optimization of Logistic Regression and Random Forest for Hoax News Detection Using TF-IDF Text Representation,” vol. 4, no. 8, pp. 381–392, 2024.

#### **LAMPIRAN**

Tabel laporan pembagian tugas

<b>Nama</b>	<b>Deskripsi Tugas</b>
<b>Samuel Leonardo Nainggolan</b>	<ol style="list-style-type: none"><li>1. Pencarian Data ke lokasi</li><li>2. Preprocessing data</li></ol>
<b>Marselino Tambunan</b>	<ol style="list-style-type: none"><li>1. Pencarian Data ke lokasi</li><li>2. Menentukan judul topik penelitian</li><li>3. Mencari jurnal sesuai kriteria</li></ol>
<b>Eduward Gilbert Simanjuntak</b>	<ol style="list-style-type: none"><li>1. Pencarian Data ke lokasi</li><li>2. Preprocessing data</li></ol>
<b>Franky Hamonangan Sirait</b>	<ol style="list-style-type: none"><li>1. Pencarian Data ke lokasi</li><li>2. Mencari jurnal sesuai kriteria</li></ol>

<b>Nathanael T. J. Tampubolon</b>	<ol style="list-style-type: none"> <li>1. Pencarian Data ke lokasi</li> <li>2. Preprocessing data</li> </ol>
<b>Whisnu Abraham Luciano Saragih</b>	<ol style="list-style-type: none"> <li>1. Pencarian Data ke lokasi</li> <li>2. Menentukan judul topik penelitian</li> <li>3. Preprocessing data</li> </ol>