

Analisis Linearitas Pendidikan terhadap Pekerjaan di Sumatera Utara

11423001 Samuel Leonardo Nainggolan

11423004 Marselino Tambunan

11423018 Eduward Gilbert Simanjuntak

11423042 Franky Hamonangan Sirait

11423044 Nathanael T. J. Tampubolon

11423045 Whisnu Abraham Luciano Saragih

Kelompok 4

Teknologi Rekayasa Perangkat Lunak



Start Slide



02

03

01

TOPIK

- LATAR BELAKANG
- DESKRIPSI DATASET
- EKSPLORATORY DATA ANALYSIS
- DATA CLEANING DAN PERBAIKAN DATA
- HASIL KINERJA MODEL
- DASHBOARD VISUALIZATION

Latar Belakang

Perkembangan ilmu data atau Data Science memungkinkan analisis hubungan antara pendidikan dan pekerjaan dilakukan secara lebih objektif dan terukur. Melalui pendekatan statistik, visualisasi data, serta pemodelan berbasis Machine Learning, pola hubungan dan tingkat kesesuaian antara bidang pendidikan dan jenis pekerjaan dapat dianalisis secara kuantitatif.

Oleh karena itu, penelitian ini bertujuan untuk menganalisis linearitas dan tingkat kesesuaian antara latar belakang pendidikan dengan jenis pekerjaan penduduk di Provinsi Sumatera Utara. Analisis dilakukan menggunakan pendekatan Data Analytics dan Machine Learning terhadap 2.000 data responden kuesioner. Hasil penelitian ini diharapkan dapat memberikan gambaran apakah mayoritas penduduk di Sumatera Utara telah bekerja sesuai dengan bidang pendidikan yang ditempuh atau sebaliknya, sehingga dapat menjadi bahan pertimbangan dalam perencanaan pendidikan dan kebijakan ketenagakerjaan.

Tujuan :

- Mengetahui distribusi tingkat pendidikan dan bidang/jurusan
- Mengetahui distribusi jenis pekerjaan
- Menganalisis hubungan dan linearitas antara tingkat pendidikan serta bidang pendidikan terhadap jenis pekerjaan
- Membangun model Machine Learning untuk memprediksi linearitas pekerjaan berdasarkan tingkat pendidikan.

Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini merupakan data primer yang diperoleh melalui penyebaran kuesioner daring menggunakan Google Form yang disebarluaskan kepada responden yang berdomisili di Provinsi Sumatera Utara dengan tujuan memperoleh informasi terkait latar belakang pendidikan, bidang pekerjaan, serta karakteristik demografis responden. Total data yang berhasil dikumpulkan sebanyak 2043 responden dengan 14 variabel.

Jumlah Kolom: 14

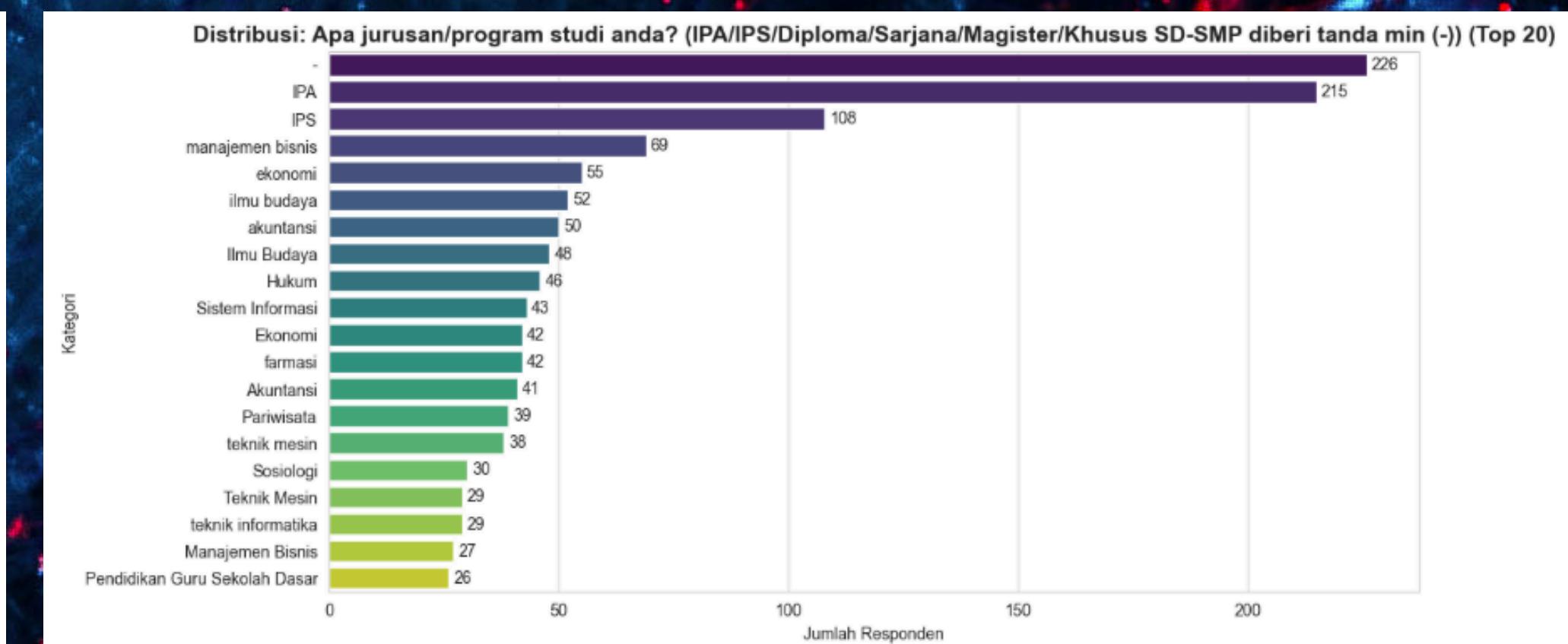
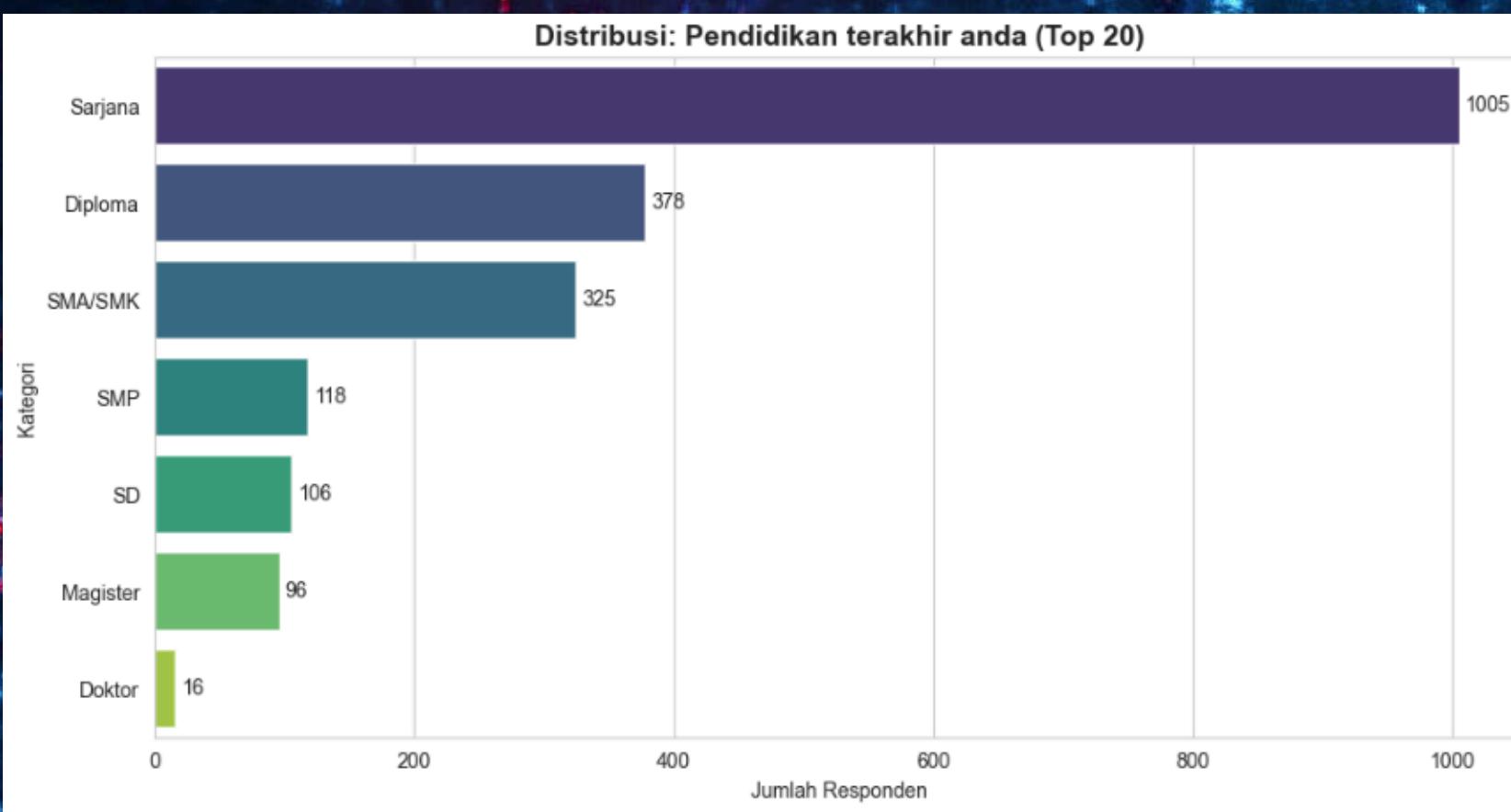
Jumlah data: 2043 responden

NO.	Attribute / Variable	Type Data
1	Usia	Numerik
2	Jenis Kelamin	Categorical
3	Kabupaten/Kota Domisili	Categorical
4	Status Pernikahan	Categorical
5	Pendidikan Terakhir	Categorical
6	Jurusan Pendidikan	Categorical
7	Status Pekerjaan	Categorical
8	Bidang Pekerjaan	Categorical
9	Lama Bekerja	Ordinal
10	Pekerjaan Sebelumnya	Categorical
11	Kesesuaian Pekerjaan Sebelumnya	Categorical
12	Penghasilan Sebelumnya	Ordinal
13	Penghasilan Saat ini	Ordinal
14	Frekuensi Penggunaan Ilmu	Ordinal

Exploratory Data Analysis



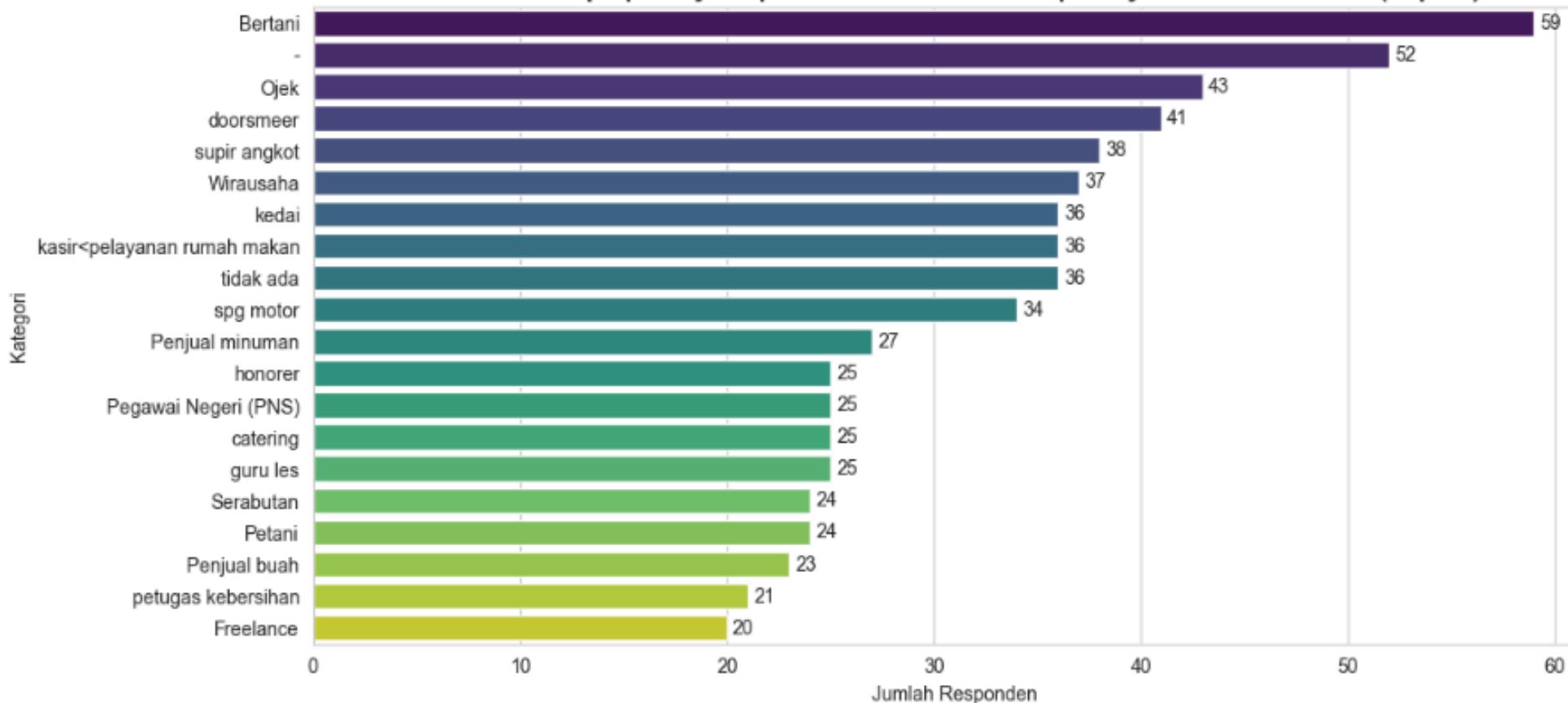
Tahap EDA bertujuan untuk memperoleh pemahaman awal terhadap karakteristik, pola, dan distribusi data sebelum dilakukan analisis lanjutan.



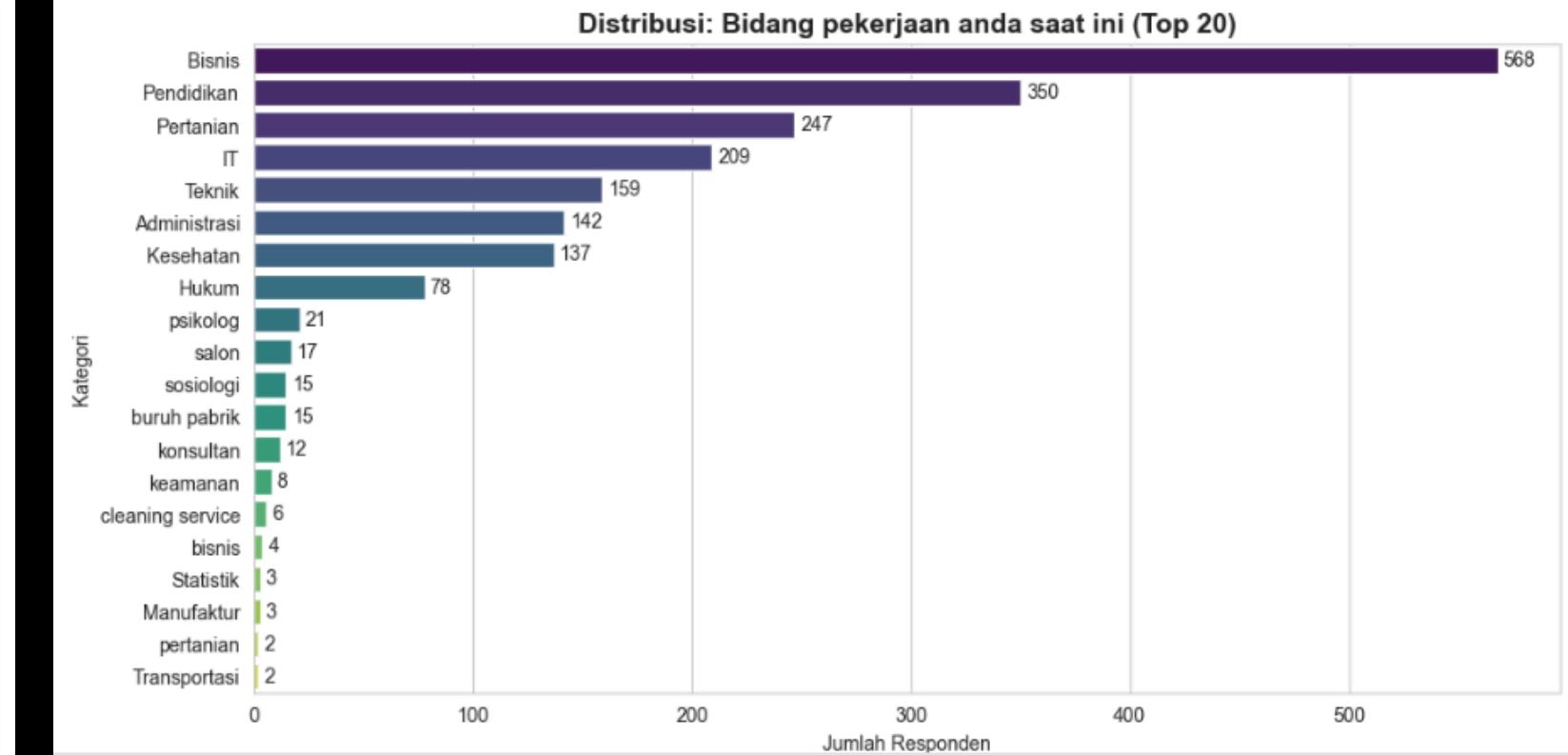
Exploratory Data Analysis



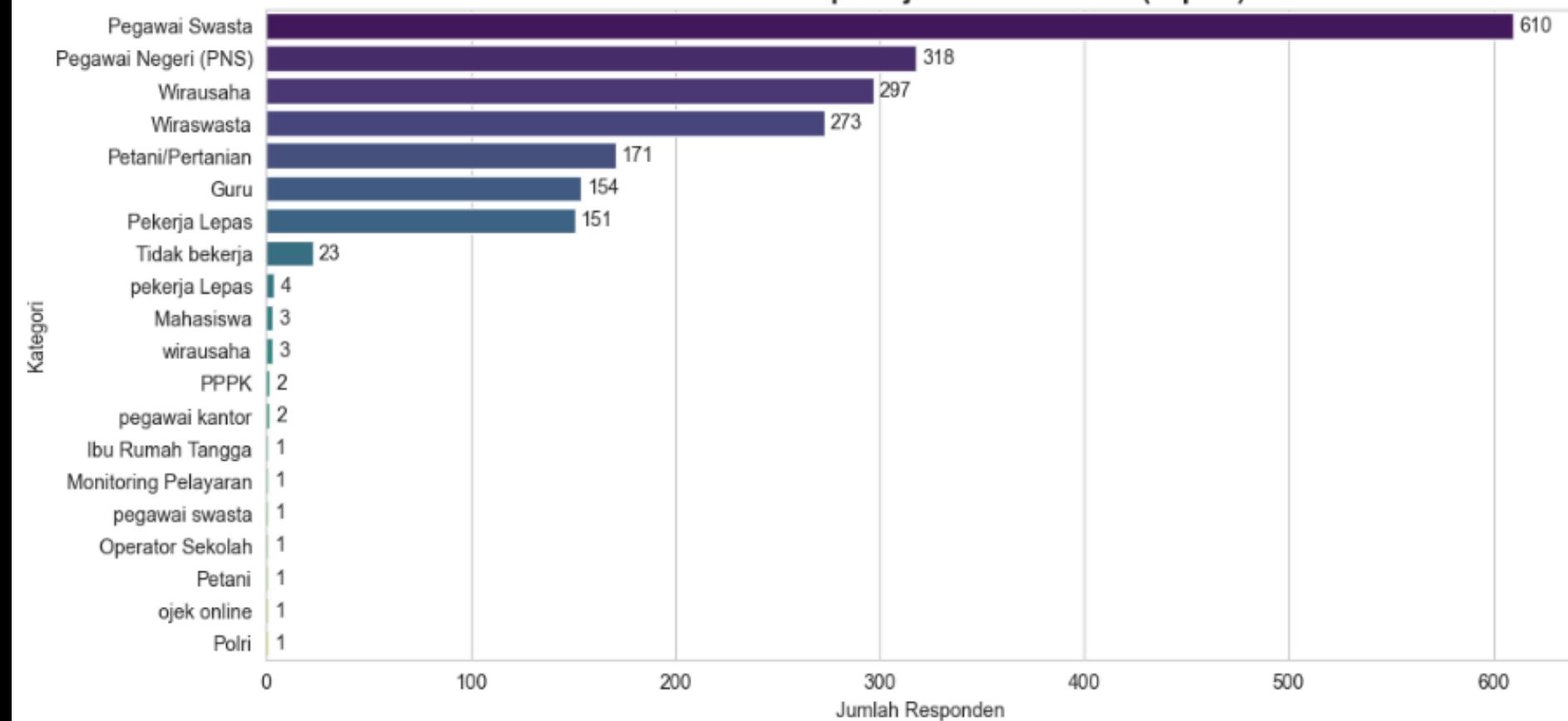
Distribusi: Apa pekerjaan pertama anda sebelum pekerjaan anda saat ini? (Top 20)



Distribusi: Bidang pekerjaan anda saat ini (Top 20)



Distribusi: Status pekerjaan anda saat ini (Top 20)



Data Cleaning dan Perbaikan Data



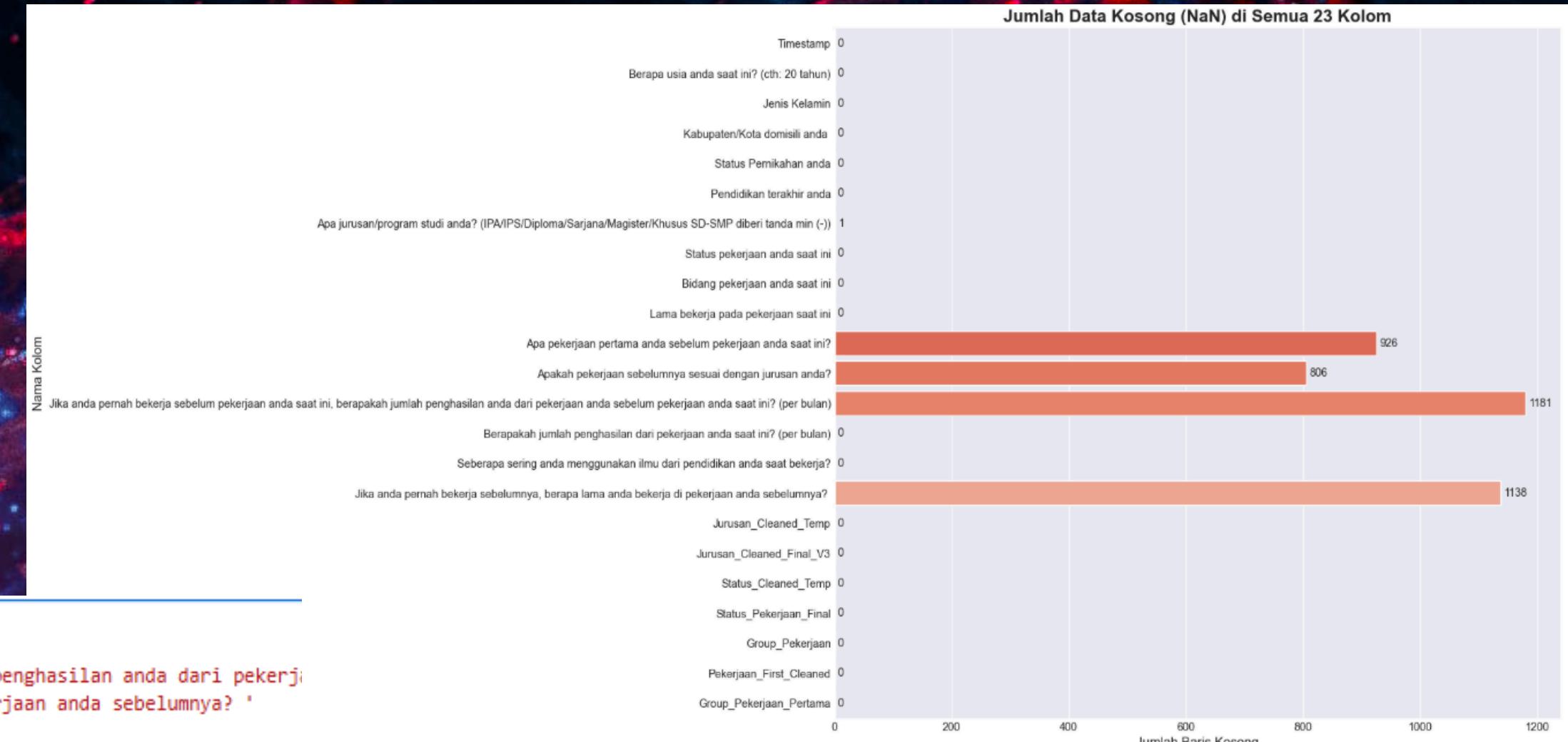
Pada data cleaning, kami menggunakan metode Missing Values yaitu proses penanganan data yang tidak lengkap atau memiliki nilai kosong. Serta mapping (membuat batasan) pada jurusan, pekerjaan serta fitur penting lainnya

```
_pekerjaan_pertama = 'Apa pekerjaan pertama anda sebelum pekerjaan anda saat ini?'
_linearitas_lama = 'Apakah pekerjaan sebelumnya sesuai dengan jurusan anda?'
_gaji_lama = 'Jika anda pernah bekerja sebelum pekerjaan anda saat ini, berapakah jumlah penghasilan anda dari pekerjaan anda sebelumnya?'
_lama_kerja_lama = 'Jika anda pernah bekerja sebelumnya, berapa lama anda bekerja di pekerjaan anda sebelumnya? '

col_pekerjaan_pertama] = df[col_pekerjaan_pertama].fillna('Belum pernah bekerja')
col_linearitas_lama] = df[col_linearitas_lama].fillna('Belum pernah bekerja')

col_gaji_lama] = df[col_gaji_lama].fillna('0')
col_lama_kerja_lama] = df[col_lama_kerja_lama].fillna('0')

nt("✓ Jumlah Data Kosong setelah pengisian spesifik:", df.isnull().sum().sum())
nt("\nCek sampel hasil pengisian:")
nt(df[[col_pekerjaan_pertama, col_gaji_lama]].head())
```



Data Cleaning dan Perbaikan Data

1	Berapa usia anda saat ini? (cth: 20 tahun)	Jenis Kelamin	Kabupaten/Kota domisili anda	Status Pernikahan anda	Pendidikan terakhir anda	Apa jurusan/program studi anda? (IPA/IPS/D)	Status pekerjaan anda saat ini
51	53	Pria	Deli Serdang	Menikah	Diploma	Perpajakan	Pegawai Negeri (PNS)
52	24	Pria	Tebing Tinggi	Belum Menikah	SMA/SMK		Wirausaha

Jumlah Baris Data: 2044 (Harusnya tidak berkurang)
Data anomali telah dihapus. Sisa jumlah baris: 2043

Distribusi Bidang Pekerjaan (Setelah '--' Dihapus):

Group_Pekerjaan	count
Bisnis	577
Pendidikan	353
Pertanian	252
IT	215
Teknik	163
Kesehatan	160
Administrasi	142
Hukum	80
Salon	17
Buruh Pabrik	16
Sosiologi	15
Konsultan	13
Keamanan	9
Cleaning Service	6
Manufaktur	3
Ibu Rumah Tangga	2
Transportasi	2
Percetakan	2
Perkantoran	2
Wiraswasta	1
Agama	1
Pelayaran	1
Tata Ruang Dan Pengairan	1
Transportasi Umum	1
Produksi	1
Pendeta	1
Tukang Kebersihan Di Pabrik	1
Sosial	1
Perikanan	1
Laundry	1
Tidak Bekerja	1
Pemerintahan	1
Peminjaman	1
	1
Name: count, dtype: int64	

===== DATA CLEANING & PERBAIKAN =====

Modus jurusan valid: IPA

Distribusi Jurusan Setelah Perbaikan (Anomali & Null diganti Modus):

Jurusan_Cleaned_Final_V3

IPA	257
Teknologi Informasi/Komputer	236
-	226
Teknologi Industri/Mesin/Elektro	152
Akuntansi	122
Ekonomi	110
IPS	109
Manajemen Bisnis	107
Ilmu Budaya/Sastra	100
Hukum	71

Name: count, dtype: int64

Proses data cleaning pada penelitian ini dilakukan untuk memastikan bahwa data kuesioner yang digunakan berada dalam kondisi bersih, konsisten, dan siap dianalisis, sehingga hasil analisis dan pemodelan yang dihasilkan bersifat akurat.

Preprocessing Data

Mengubah data mentah menjadi format yang mudah dipahami oleh algoritma.

1. Standardisasi & Grouping (Text Cleaning)
2. Encoding
3. Penentuan Label Target (Rule-Based Labeling)

```
def standardize_jurusan_v3(jurusan):
    groups = {
        # --- TEKNOLOGI ---
        'Teknologi Rekayasa Perangkat Lunak': ['teknologi rekayasa', 'perangkat lunak', 'trpl', 'rekayasa perangkat'],
        'Teknologi Informasi/Komputer': ['teknologi informasi', 'informatika', 'sistem informasi', 'komputer', 'ti', 'it', 'rpl', 'coding', 'software']

        # --- PENDIDIKAN ---
        'Pendidikan Guru Sekolah Dasar (PGSD)': ['pgsd', 'guru sekolah dasar'],
        'Pendidikan Bahasa': ['pendidikan bahasa', 'sastra inggris', 'bahasa inggris', 'bahasa indonesia', 'sastra indonesia'],
        'Pendidikan Agama': ['pendidikan agama', 'teologi', 'theologia', 'pendidikan kristen', 'pendidikan islam'],
        'Pendidikan MIPA': ['pendidikan biologi', 'pendidikan kim', 'pendidikan fis', 'pendidikan mat', 'pendidikan ipa', 'keguruan (biologi)'],
        'Pendidikan Olahraga': ['pendidikan olahraga', 'kepelatihan olahraga', 'ilmu keolahragaan'],
        'Pendidikan (Umum/Lainnya)': ['pendidikan', 'keguruan', 'guru', 'kependidikan'],
    }
```

```
def kelompokkan_pekerjaan(nilai):
    if pd.isna(nilai) or str(nilai).strip() == '':
        return 'Tidak Diisi'
    val_lower = str(nilai).lower().strip()
    val_asli_rapi = str(nilai).strip().title()

    # 1. IT
    if any(keyword in val_lower for keyword in ['it', 'informatika', 'software', 'program', 'developer', 'sistem', 'cyber', 'data', 'network', 'web', 'statistika']):
        return 'IT'
    # 2. Kesehatan
    elif any(keyword in val_lower for keyword in ['sehat', 'dokter', 'perawat', 'bidan', 'apotek', 'klinik', 'medis', 'rumahsakit', 'farmasi', 'kesehatan', 'psikolog']):
        return 'Kesehatan'
    # 3. Pendidikan
```

```
#status pekerjaan yang doduble
kolom_status = 'Status pekerjaan anda saat ini'
df['Status_Cleaned_Temp'] = df[kolom_status].fillna('nilai_hilang').astype(str).str.lower().str.strip()

def standardize_status_pilihan_berganda(status):
    if status == 'nilai_hilang' or status == 'nan':
        return np.nan

    # 1. CEK KELOMPOK GURU (Target: "Guru")
    # Keyword: guru, dosen, pengajar, tentor, honorer sekolah
    if any(kw in status for kw in ['guru', 'dosen', 'pengajar', 'tentor', 'pendidik', 'sekolah']):
        return 'Guru'

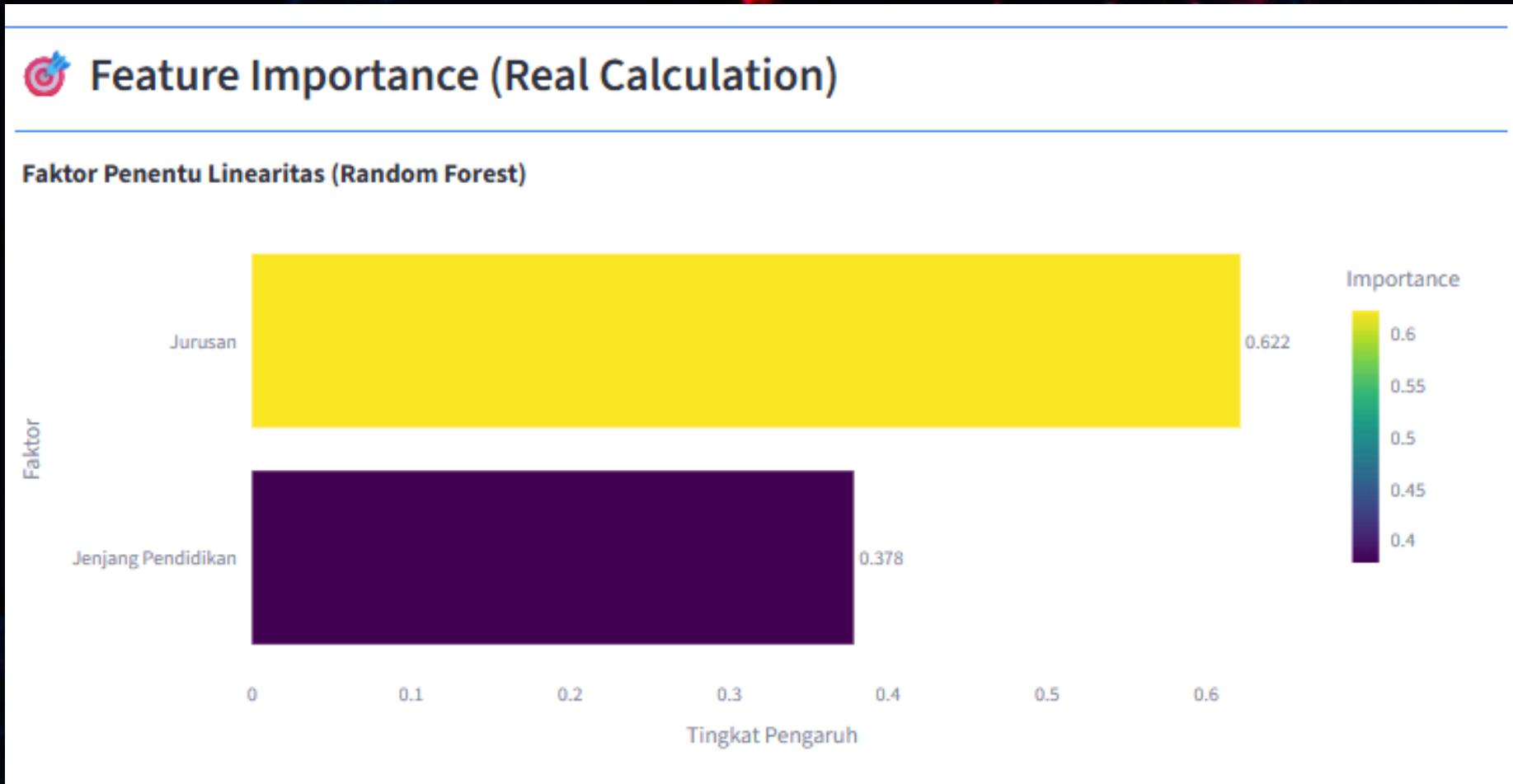
    # 2. CEK KELOMPOK PNS (Target: "Pegawai Negeri (PNS)")
    # Keyword: pns, ASN, PPPK, CPNS, pegawai negeri
    if any(kw in status for kw in ['pns', 'ASN', 'pppk', 'cpns', 'pegawai negeri', 'nusantara sehat']):
        return 'Pegawai Negeri (PNS)'
```

Preprocessing Data

```
col_pend = 'Pendidikan terakhir anda'  
map_pend = {  
    'Tidak sekolah': 0,  
    'SD': 1,  
    'SMP': 2,  
    'SMA/SMK': 3,  
    'Diploma': 4,  
    'Sarjana': 5,  
    'Magister': 6,  
    'Doktor': 7,  
    'Yang lain': 0  
}  
  
[  
  
    'Jurusan_Cleaned_Final_V3',  
    'Status_Pekerjaan_Final',  
    'Group_Pekerjaan',  
    'Group_Pekerjaan_Pertama'  
]  
  
cols_to_encode = [col for col in cols_to_encode if col in df.columns]  
le = LabelEncoder()  
  
for col in cols_to_encode:  
    new_col_name = col + '_Encoded_Label'  
    df[new_col_name] = le.fit_transform(df[col].astype(str))  
    print(f"    -> {col} diubah menjadi angka otomatis.")  
  
print("✅ [3/3] Label Encoding Variabel Lain Selesai.")
```

```
aturan_linear_v3 = {  
    # --- 1. BIDANG TEKNOLOGI & KOMPUTER ---  
    'Teknologi Informasi/Komputer': [  
        'teknologi informasi', 'komputer', 'software', 'ti', 'it', 'program',  
        'data', 'jaringan', 'teknisi', 'sistem', 'web', 'app', 'android', |  
        'digital', 'cyber', 'admin server', 'support', 'edp', 'coding',  
        'guru', 'pengajar', 'dosen', 'pendidikan', 'sekolah', 'laboran'  
    ],  
    'Teknologi Rekayasa Perangkat Lunak': [  
        'software', 'perangkat lunak', 'programmer', 'developer', 'it',  
        'teknologi', 'coding', 'web', 'app', 'system',  
        'guru', 'pengajar', 'dosen', 'pendidikan'  
    ],  
  
    # ======  
    # FUNGSI PENENTU OTOMATIS  
    # ======  
    def tentukan_linearitas_objektif_v3(baris):  
        jurusan = baris['Jurusan_Cleaned_Final_V3']  
        pekerjaan_raw = str(baris['Bidang pekerjaan anda saat ini']).lower()  
  
        # Sekarang kita memanggil variabel 'aturan_linear_v3' yang sudah diupdate  
        if jurusan in aturan_linear_v3:  
            kata_kunci_cocok = aturan_linear_v3[jurusan]  
  
            # LOGIKA: Cek jika salah satu kata kunci muncul di teks pekerjaan  
            if any(kunci in pekerjaan_raw for kunci in kata_kunci_cocok):  
                return 1 # LINEAR  
  
            return 0 # TIDAK LINEAR  
  
        # ======  
        # EKSEKUSI  
        # ======  
        print("💡 Sedang memproses ulang Linearitas dengan Aturan V3 (Logic Final)...")  
        df['Linearitas_System_Encoded'] = df.apply(tentukan_linearitas_objektif_v3, axis=1)  
  
        print("\n✅ SELESAI! Kolom 'Linearitas_System_Encoded' telah diperbarui.")  
        print("Verifikasi Sampel: Akuntansi yang bekerja di Bisnis/Dagang:")
```

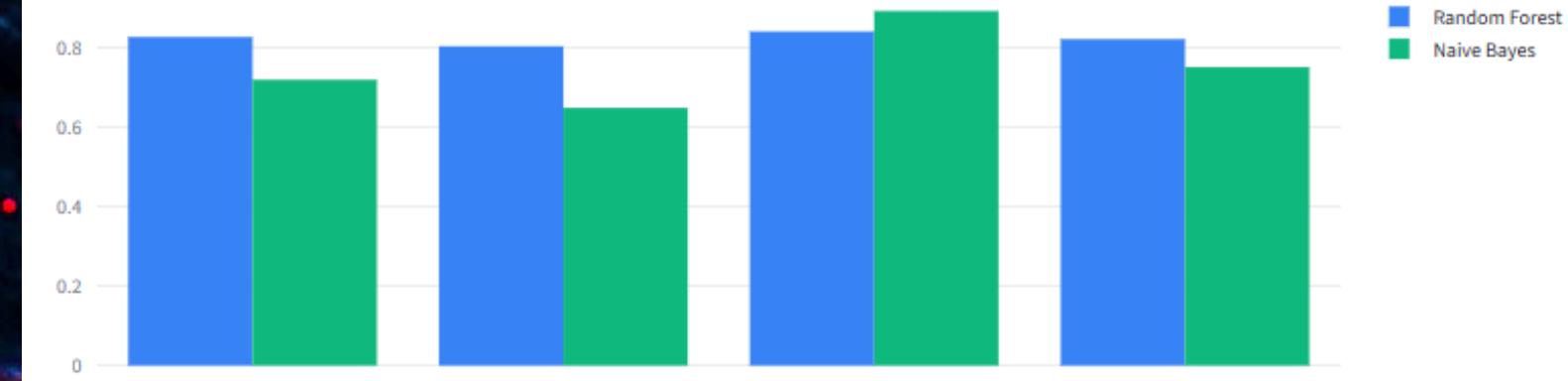
Hasil Kinerja Model



⌚ Analisis Model Machine Learning

📊 Performa Model (Real-time)

Perbandingan Performa Model ML (Real Calculation)



Random Forest
Naive Bayes

Temuan ini mengindikasikan bahwa meskipun peningkatan jenjang pendidikan berperan dalam peluang kerja, kecocokan antara jurusan dan jenis pekerjaan tetap menjadi faktor kunci dalam menentukan apakah seseorang bekerja secara linear atau tidak.

Pada hampir seluruh matrik evaluasi—akurasi, presisi, recall, dan F1-score—Random Forest menunjukkan kinerja yang lebih stabil dan seimbang, yang mengindikasikan kemampuannya dalam menangkap pola data yang lebih kompleks.

Dashboard Interaktif

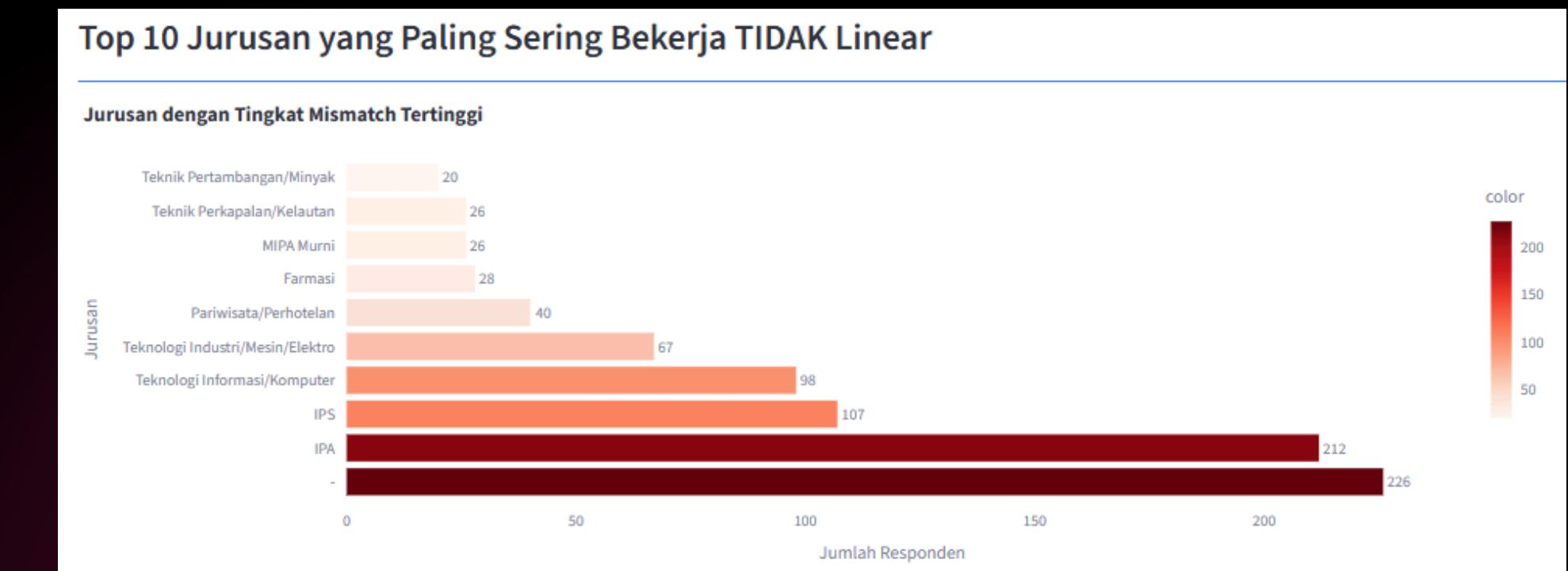
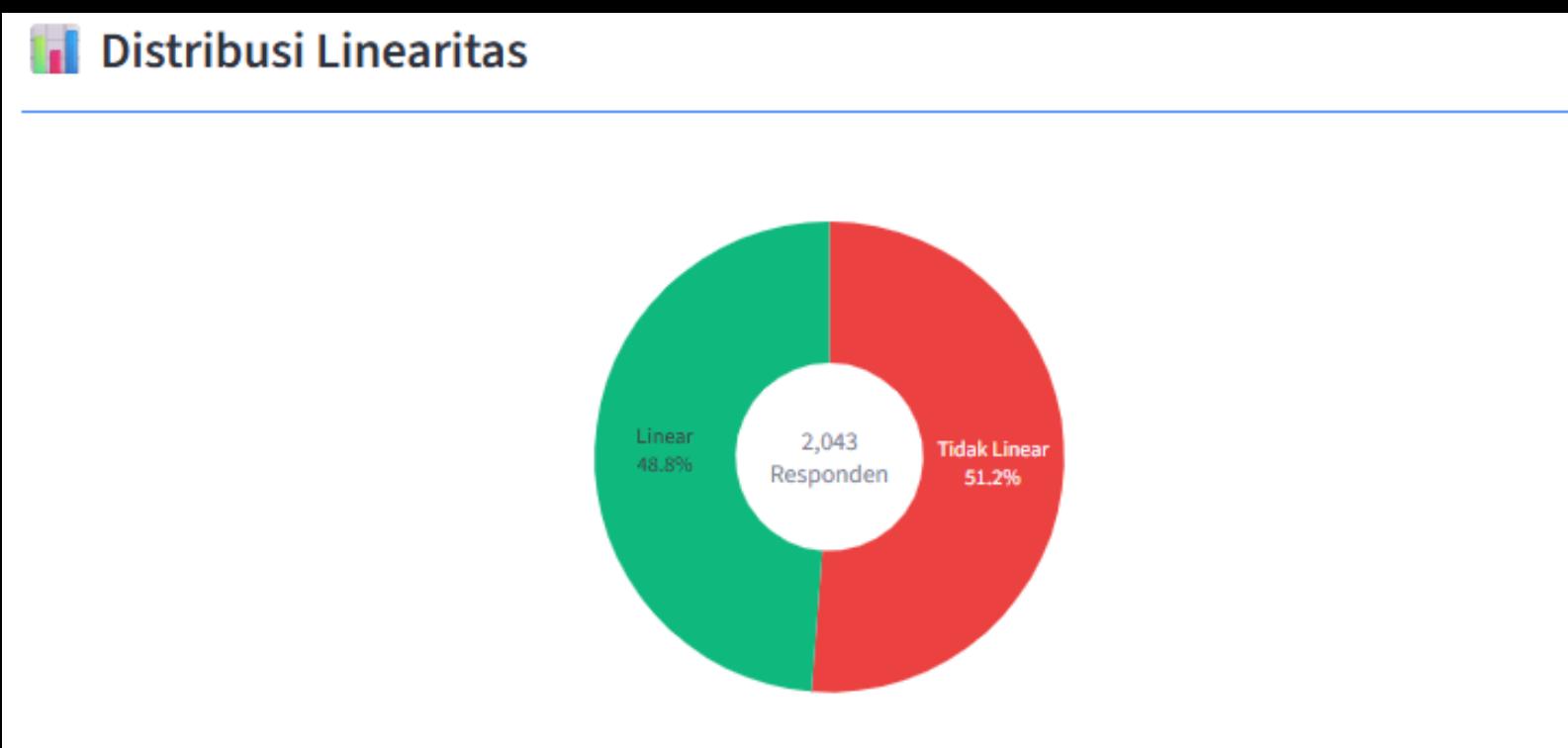
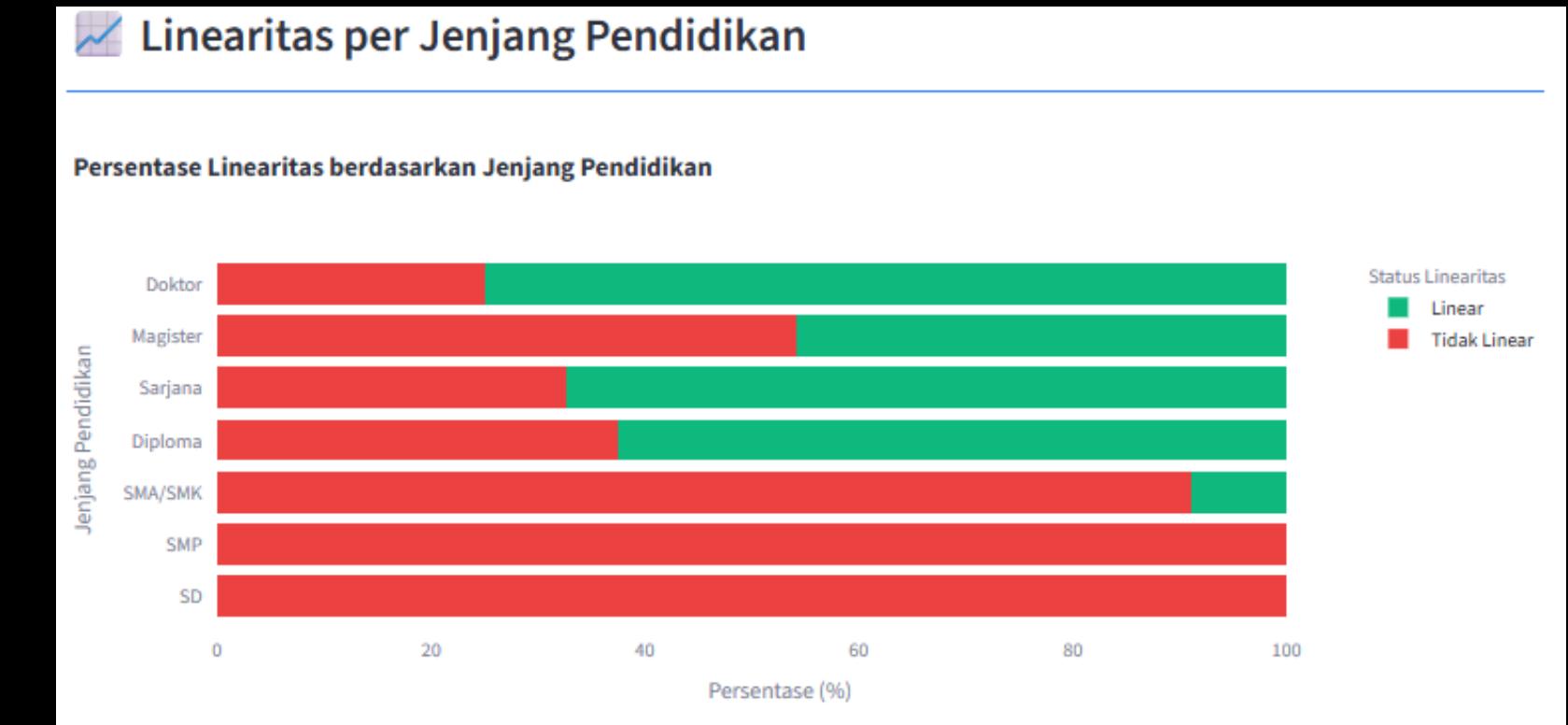
Analisis terhadap 2,043 responden di Sumatera Utara menunjukkan bahwa 51.2% bekerja di luar bidang pendidikannya. Jurusan pendidikan adalah faktor penentu utama terhadap keputusan karir

Menampilkan data 21-40 dari 997 kasus

Pendidikan_Awal	Jurusan	Status_Pekerjaan	Bidang_Pekerjaan	Status_Linearitas	Rentang_Gaji	Jenjang_Pendidikan	Gender	Status_Pernikahan	Pengalaman	Pekerjaan_Pertama
74	SMA/SMK	IPA	Pegawai Swasta	Administrasi	SESUAI	3-5 Juta	SMA/SMK	Pria	Belum Menikah	3-5 Tahun
79	SMA/SMK	IPA	Wiraswasta	Kesehatan	SESUAI	5-10 Juta	SMA/SMK	Pria	Belum Menikah	3-5 Tahun
80	SMA/SMK	IPA	Guru	Kesehatan	SESUAI	1-3 Juta	SMA/SMK	Pria	Menikah	3-5 Tahun
81	Sarjana	Pendidikan Guru Sekolah Dasar (PGSD) Pegawai Negeri (PNS)	Pendidikan	SESUAI	3-5 Juta	Sarjana	Pria	Menikah	3-5 Tahun	Guru / Pengajar
99	Sarjana	Manajemen	Pegawai Swasta	Pendidikan	SESUAI	10-15 Juta	Sarjana	Pria	Menikah	3-5 Tahun
106	Sarjana	Teknik Lainnya	Pegawai Swasta	Teknik	SESUAI	< 1 Juta	Sarjana	Pria	Belum Menikah	3-5 Tahun
107	SMA/SMK	IPA	Wirausaha	Administrasi	SESUAI	10-15 Juta	SMA/SMK	Pria	Belum Menikah	3-5 Tahun

Menampilkan data 1-20 dari 1046 kasus

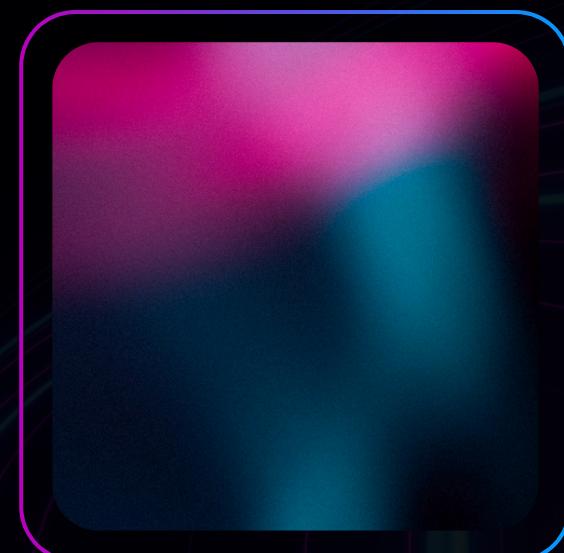
Pendidikan_Awal	Jurusan	Status_Pekerjaan	Bidang_Pekerjaan	Status_Linearitas	Rentang_Gaji	Jenjang_Pendidikan	Gender	Status_Pernikahan	Pengalaman	Pekerjaan_Pertama
0	Sarjana	Teknologi Informasi/Komputer	Ibu Rumah Tangga	Ibu Rumah Tangga	TIDAK SESUAI	1-3 Juta	Sarjana	Pria	Menikah	> 5 Tahun
3	Diploma	Teknologi Industri/Mesin/Elektro	Wiraswasta	Bisnis	TIDAK SESUAI	5-10 Juta	Diploma	Pria	Menikah	> 5 Tahun
4	SMA/SMK	Stm Negeri Balige	Wiraswasta	Bisnis	TIDAK SESUAI	5-10 Juta	SMA/SMK	Pria	Menikah	> 5 Tahun
5	SMA/SMK	IPA	Wiraswasta	Bisnis	TIDAK SESUAI	1-3 Juta	SMA/SMK	Pria	Menikah	> 5 Tahun
6	Diploma	Teknologi Informasi/Komputer	Wirausaha	Bisnis	TIDAK SESUAI	1-3 Juta	Diploma	Pria	Menikah	> 5 Tahun
8	Diploma	Teknologi Informasi/Komputer	Wirausaha	Wiraswasta	TIDAK SESUAI	1-3 Juta	Diploma	Pria	Menikah	> 5 Tahun





Thank You

FOR YOUR ATTENTION



**Kelompok 04
Teknologi Rekayasa Perangkat Lunak**