Teknik Metode Pemrosesan Data

Mobilitas Vertikal IFLS 2007 - 2014

1. LATAR BELAKANG DATA

IFLS (The Indonesia Family Life Survey) adalah survei longitudinal tentang kondisi sosial ekonomi dan kesehatan yang dilakukan oleh RAND Corporation bekerja sama dengan Universitas Gadjah Mada. Dokumen ini menjelaskan prosedur pengolahan data yang digunakan untuk membangun dataset analitis yang sesuai untuk menganalisis hubungan antara lembaga pendidikan Islam dan hasil ekonomi yang didapat.

2. SUMBER DATA

2.1. Sumber Data Utama

Dataset : Indonesia Family Life Survey (IFLS)

Gelombang yang digunakan : 2007 (Wave 4) dan 2014 (Wave 5)

Cakupan Geografis : 13 dari 34 Provinsi (2014)

Ukuran Sampel Asli:

- IFLS 2007: ~ 13.500 rumah tangga, ~ 45.000 individu
- IFLS 2014: ~ 16.200 rumah tangga, ~ 50.000 individu

2.2. Modul yang Diambil

Berikut adalah modul IFLS yang digunakan untuk konstruksi data:

Kode Modul	Deskripsi	Variabel yang diguankan
B3A_COV	Halaman Sampul / Demografi	ID individu (pidlink), usia, gender
B3A_DL1	Buku Pendidikan 1	Riwayat pendidikan tertinggi, riwayat pernah pesantren
B3A_DL2	Buku Pendidikan 2	Tipe sekolah, riwayat sekolah
B3A_DL4	Buku Pendidikan 4	Tahun masuk, lulus, dan pengulangan kelas
B3A_TK3	Buku Pekerjaan 3	Tipe pekerjaan, status pekerjaan
BK_AR1	Ekonomi Rumah Tangga	Pendapatan Tahunan
BK_SC1	Karakteristik Komunitas	Desa/Kota, Kode Geografi

2.3. Data Tambahan

Indeks Harga Konsumen (IHK): Diperoleh dari Badan Pusat Statistik (BPS) Indonesia untuk tujuan deflasi pendapatan.

- Tahun Dasar : 2007 (100)
- Tahun Deflator : 2014 (150,47)

3. Sampel dan Populasi

3.1. Target Populasi

Sampel analisis ini berfokus pada individu usia kerja (usia 15-65 tahun) yang:

- Berpartisipasi dalam kedua gelombang IFLS 2007 dan 2014
- Memiliki catatan riwayat pendidikan yang lengkap
- Memiliki identifikasi rumah tangga yang valid untuk pengaitan orang tua

3.2. Konstruksi Sampel

Berdasar modul B3A_COV (IFLS 2014), pada awalnya terdapat 36.391 individu. Selanjutnya mengalami filter dan *merging* secara sekuensial, diantaranya:

- *merging* dengan riwayat pendidikan : 34.464 individu,
- *merging* dengan data geografi 2007: 21.979 individu,
- *merging* dengan data pendapatan: 10.878 individu, dan
- *merging* dengan data pekerjaan: 10.878 individu. Dan

Dari sini didapatkan *attrition rate* sebesar 70,1% dari data awal tahun 2014. Selain itu pada kolom latar belakang pendidikan orang tua, hanya terdapat ~ 922 individu yang dapat di *tracing*, sedangkan sisanya dibiarkan kosong, dengan tujuan analisis deskriptif.

4. METODE PEMROSESAN DATA

4.1. Strategi Integrasi Data

4.1.1. Identifier Unik

Disini digunakan primary key berupa pidlink (Person ID Link). Fitur ini memiliki karakteristik sebagai berikut pada IFLS.

- Konsisten di seluruh gelombang IFLS
- Format: Kode alfanumerik 9 digit
- Contoh: “001220001”

Selain itu, juga digunakan hhid (household ID), dengan karakteristik sebagai berikut pada IFLS.

- Format: Kode 9 karakter
- Digunakan untuk analisis hubungan keluarga

4.1.2. Prosedur Merge

Penggabungan data dilakukan menggunakan beberapa operasi yang berbeda, outer-join pada kasus dimana data tidak *overlapping*, left-join untuk mempertahankan ukuran sampel yang maksimal, dan inner-join untuk memastikan ketersediaan variabel hasil.

```
Base Dataset (B3A_COV)
  ← LEFT JOIN Education (DL1/DL2/DL4)
  ← INNER JOIN Geography (SC1)
  ← INNER JOIN Income (AR1)
  ← LEFT JOIN Employment (TK3)
```

4.2. Penganganan Data Hilang (Missing Value)

- Analisis kasus lengkap untuk variabel utama; variabel indikator untuk kelengkapan data tambahan yang hilang.
- Imputasi: Tidak dilakukan. Nilai yang hilang dikodekan sebagai NaN (Python/Pandas) atau . (Stata).
- Alasan: Mengingat sifat longitudinal dan potensi atrisi non-acak, imputasi dianggap tidak tepat.

4.3. Penanganan Duplikasi

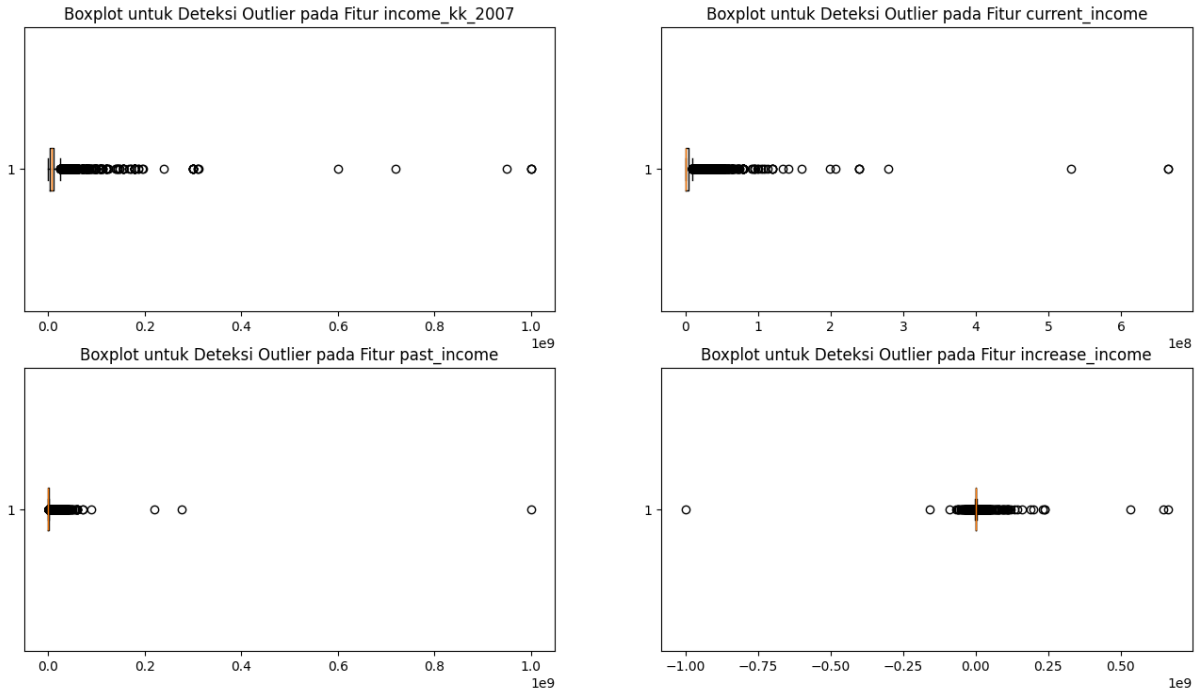
Nilai pidlink yang duplikat teridentifikasi dalam modul ketenagakerjaan dan pendidikan akibat adanya beberapa catatan pekerjaan/sekolah per individu. Hal ini dapat ditangani dengan mengambil nilai dari indeks terakhir. Catatan terbaru dipertahankan, dengan asumsi bahwa catatan tersebut mewakili status terkini.

```
df = df.drop_duplicates(subset=['pidlink'], keep='last')
```

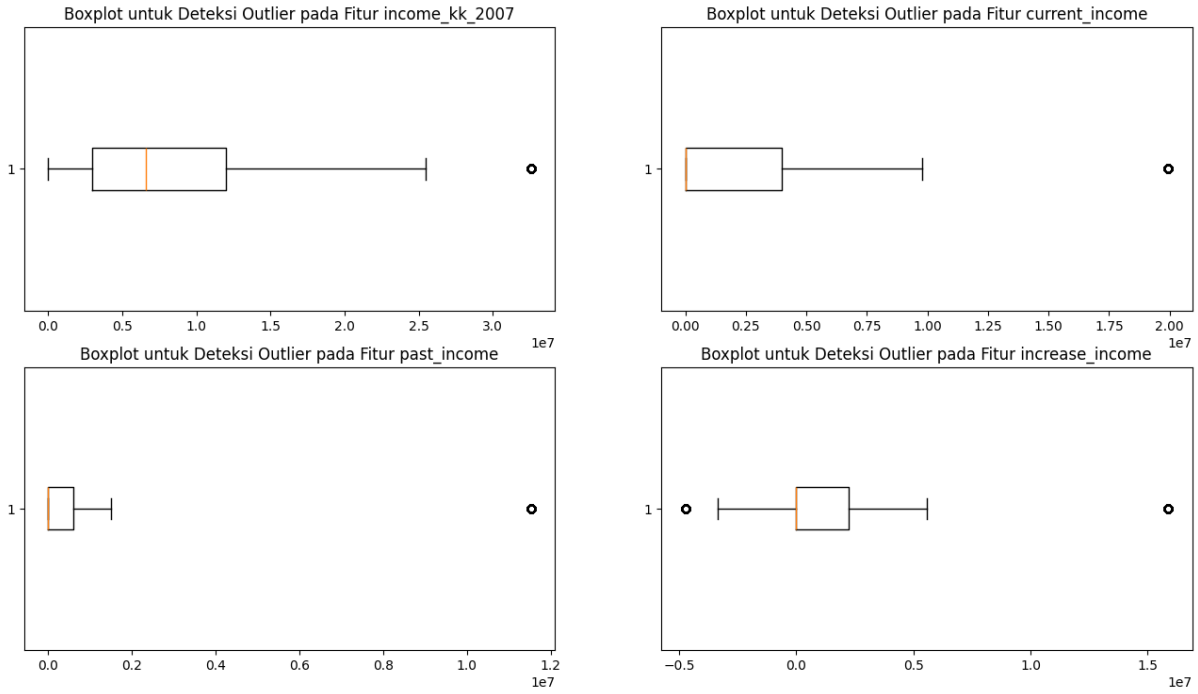
4.4. Penanganan Outlier

Dalam analisis ini outlier dideteksi menggunakan metode IQR (*interquartile range*), dimana nilai yang berada di luar rentang $Q_1 - 1{,}5 \times IQR$ dan $Q_1 + 1{,}5 \times IQR$ dengan

$IQR = Q3 - Q1$ dikategorikan sebagai outlier. Setelah outlier dideteksi, beberapa penanganan yang dilakukan adalah deteksi validitas data. Pertama adalah validitas data. Misalnya dalam semua fitur, seharusnya semuanya bernilai positif, jika negatif, akan kami lihat apakah memungkinkan kesalahan penginputan tanda minus pada data, jika tidak nilainya akan kami interpolasi dengan data tahun sebelum atau setelahnya. Kedua adalah menghapus data. Penghapusan data kami lakukan secara hati-hati. Teknik yang kami lakukan adalah jika semua fitur dalam satu data merupakan outlier, maka data tersebut berpotensi dihapus. Ketiga adalah *winsorizing*, yaitu mengganti nilai ekstrem dengan nilai batas tertentu. Penelitian ini akan menggunakan persentil-5 dan persentil-95 sebagai batas bawah dan atas.



Adapun fitur yang dilakukan penanganan outlier adalah fitur yang berhubungan dengan nilai ekonomi, diantaranya `income_kk_2007`, `current_income`, `past_income`, dan `increase_income`. Berikut adalah persebaran data setelah penanganan outlier.



5. FEATURE ENGINEERING (KONSTRUKSI VARIABEL)

5.1. Variabel Tingkat Pendidikan

5.1.1. Lama Menempun Sekolah (years of schooling)

Durasi pendidikan disusun secara terpisah untuk setiap tingkat menggunakan algoritma berikut:

Sekolah Dasar

```
long_year_sd = 6 + number_grade_fail
```

Dimana, number_grade_fail adalah hasil perhitungan kelas yang diulang.

Sekolah Menengah Pertama dan Atas (SMP atau SMA)

```
long_year_smp/sma = 3 + number_grade_fail
```

Universitas

```
long_year_univ = |graduation_year - entry_year|
```

Total tahun menempuh akademik

```
long_year_acad = long_year_sd + long_year_smp +
                 long_year_sma + long_year_univ
```

5.1.2. Indikator Pendidikan Islam

Sekolah diklasifikasikan sebagai lembaga Islam berdasarkan variabel dl11:

Definisi Sekolah Islam:

- 2: Sekolah agama negeri (Madrasah Negeri)
- 4: Sekolah Islam swasta (Madrasah Swasta)

Untuk gambaran lebih jelas, dapat dilihat dari konstruksi indikator berikut:

```
islamic_school = ['2:Public religious', '4:Private islam']
df['sd_islam'] = np.where(df['dl11'].isin(islamic_school),
1, 0)
```

Proses ini direplikasi untuk setiap level pendidikan (SMP, SMA, Universitas).

Lama Pendidikan Islam, didapatkan dengan mengalikan variabel biner sekolah islam dengan lama studi pada setiap jenjang.

```
long_year_acad_islam_educ = (sd_islam × long_year_sd) +
                            (smp_islam × long_year_smp)
   +
                            (sma_islam × long_year_sma)
   +
                            (univ_islam                ×
   long_year_univ)
```

5.1.3. Pendidikan Tertinggi Pesantren

Merupakan indikator biner yang didapatkan dari dl06x:

```
highest_educ_pesantren = 1 if dl06x == "1:Yes", else 0
```

5.1.4. Latar Belakang Pendidikan Islam Terpadu

Fitur islamic_educ_background merupakan fitur/variabel biner dimana bernilai benar (1) jika tingkat pendidikan tertingginya meruapakan pondok pesantren atau setidaknya pernah mengikuti pendidikan formal Islam sesuai definisi sebelumnya.

5.2. Variabel Pendapatan

5.2.1. Deflasi Pendapatan

Untuk memastikan kesesuaian antar tahun survei, pendapatan tahun 2014 disesuaikan ke harga konstan tahun 2007:

```
current_income = ar15b_2014 × (100 / 150.47)
past_income = ar15b_2007
```

di mana 150,47 adalah nilai IHK untuk tahun 2014 (dasar 2007 = 100).

5.2.2. Variabel Peningkatan Pendapatan

Didapatkan dengan mengurangi, nilai pendapatan 2014 (setelah di deflasikan) dengan nilai pendapatan 2007. Ini menunjukkan perubahan pendapatan riil antara tahun 2007 dan 2014.

5.3. Karakteristik Orang Tua

5.3.1. Identifikasi Kepala Keluarga

Kepala Keluarga di identifikasi dengan menggunakan variabel hubungan ar02b:

```
household_head = 1 if ar02b == "1:Head of household"
```

5.3.2. Variabel Pendapatan Orang Tua

Selanjutnya, identifier orang tua tersebut akan ditelusuri nilai pendapatannya, dan disimpan kedalam suatu objek yang selanjutnya akan di petakan pada setiap anggota keluarga.

```
dict_income_kk = {}
for idx in df_kk.index:
    hhid = df_kk.loc[idx, 'hhid07_9']
    if hhid not in dict_income_kk:
        dict_income_kk[hhid] = df_kk.loc[idx, 'ar15b']
    else:
        print("Double)
        dict_income_kk[hhid] += df_kk.loc[idx, 'ar15b']
```

5.3.3. Variabel Pendidikan Orang Tua

Metrik pendidikan orang tua dipetakan berdasarkan karakteristik kepala rumah tangga:

- `long_year_acad_kk_2007`: Total tahun pendidikan formal
- `long_year_acad_islam_educ_kk_2007`: Tahun di lembaga pendidikan Islam
- `islamic_educ_background_kk_2007`: Indikator pendidikan Islam biner

5.4. Variabel Pekerjaaan

5.4.1. Klasifikasi Status Pekerjaan

Disini kode status pekerjaan asli diubah menjadi kategori standar:

Kode Asli	Kategori Standar
1,2,3	Wiraswasta
4	Pegawai Pemerintah
5	Pegawai Swasta
6,7,8	Buruh
9, Missing	Missing/Tidak Bekerja

5.4.2. Jenis Pekerjaan

5.4.3. Klasifikasi Umur

Untuk tujuan analisis, kami membagi umur kedalam beberapa kategori, dengan sebaran yang diperlihatkan oleh tabel berikut.

Klasifikasi Umur	Deskripsi Kelas	Frekuensi
15 – 30	Responden berusia kurang dari 30 tahun	4349
30 – 45	Responden berusia lebih dari sama dengan 30 tahun, kurang dari 45 tahun	3120
45 – 55	Responden berusia lebih dari sama dengan 45 tahun, kurang dari 55 tahun	1789
> 55	Responden berusia lebih dari sama dengan 55 tahun	1620