RESEARCH ARTICLE • OPEN ACCESS

Propensity Score Matching Pada Pemanfaatan Data Hasil *Web Scraping* Untuk Perbaikan Statistik Resmi

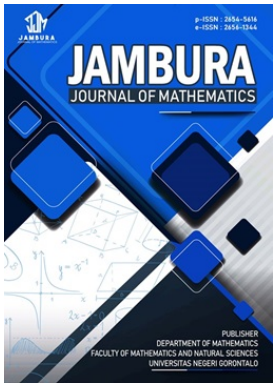
Fatimah, Hari Wijayanto, dan Farit Mochamad Afendi

Volume 6, Issue 2, Pages 226–235, August 2024

Diterima 12 Juni 2024, Direvisi 28 Juli 2024, Disetujui 2 Agustus 2024, Diterbitkan 4 Agustus 2024

To Cite this Article : F. Fatimah, H. Wijayanto, dan F. M. Afendi,"*Propensity Score Matching* Pada Pemanfaatan Data Hasil *Web Scraping* Untuk Perbaikan Statistik Resmi", *Jambura J. Math*, vol. 6, no. 2, pp. 226–235, 2024, *https://doi.org/10.37905/jjom.v6i2.26568*

JOURNAL INFO • JAMBURA JOURNAL OF MATHEMATICS

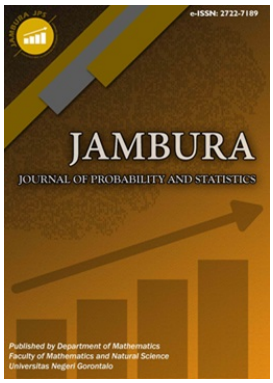


Homepage	:	http://ejurnal.ung.ac.id/index.php/jjom/index
Journal Abbreviation	:	Jambura J. Math.
Frequency	:	Biannual (February and August)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjom
Online ISSN	:	2656-1344
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjom/oai
Google Scholar ID	:	iWLjgaUAAAAJ
Email	:	info.jjom@ung.ac.id

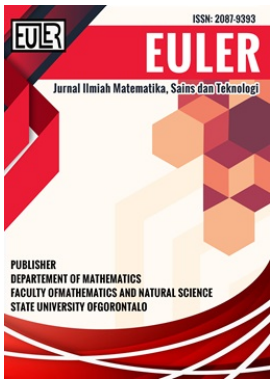
JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics

Jambura Journal of Mathematics Education



Jambura Journal of Probability and Statistics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Mathematics, Volume 6, Issue 2, Pages 226–235, August 2024 https://doi.org/10.37905/jjom.v6i2.26568

Research Article

Check for updates

Propensity Score Matching Pada Pemanfaatan Data Hasil *Web Scraping* Untuk Perbaikan Statistik Resmi

Fatimah[1,*], **Hari Wijayanto**[1], **dan Farit Mochamad Afendi**[1]

[1]*Departemen Statistika, IPB University, Bogor, Indonesia*

ARTICLE HISTORY
Diterima 12 Juni 2024
Direvisi 28 Juli 2024
Disetujui 2 Agustus 2024
Diterbitkan 4 Agustus 2024

KATA KUNCI
Big Data
Propensity Score Matching
Statistik Resmi
Tarif Kontrak Rumah
Web Scraping

KEYWORDS
Big Data
House Contract Rates
Propensity Score Matching
Official Statistics
Web Scraping

ABSTRAK. *Badan Pusat Statistik (BPS) menyambut baik tantangan untuk memanfaatkan big data. Salah satu publikasi BPS yang dapat didukung dengan menggunakan big data adalah angka inflasi yang dikumpulkan dari survei harga konsumen. Salah satu bagian dari survei harga konsumen adalah Survei HK-4 yang memuat tarif kontrak rumah. Selama ini tarif kontrak rumah yang dihasilkan BPS underestimate atau lebih rendah dari keadaan sebenarnya. Perbaikan tarif kontrak rumah dilakukan dengan matching data BPS dan web scraping situs sewa rumah menggunakan Propensity Score Matching (PSM). Data yang digunakan dalam penelitian ini meliputi DKI Jakarta, Bandung, dan Semarang pada bulan September hingga Oktober 2023. Penelitian ini bertujuan untuk mencari model matching terbaik menggunakan PSM untuk memperbaiki statistik resmi (tarif kontrak rumah) dengan menggabungkan beberapa metode pendugaan nilai propensity score dan algoritma matching. Selanjutnya hasil matching dengan model terbaik akan digunakan untuk menghitung tarif kontrak rumah terkoreksi. Hasil penelitian menunjukkan bahwa model matching terbaik secara umum menggunakan pendugaan nilai propensity score regresi logistik, algoritma nearest neighbor matching dengan pengembalian dan menggunakan rasio 1:1. Tarif kontrak terkoreksi jauh di atas tarif kontrak resmi (DKI Jakarta terkoreksi 87,27%, Bandung 316,15%, dan Semarang 60,04%). Web Scraping memungkinkan digunakan sebagai opsi untuk memperbaiki statistik resmi karena hemat biaya dan waktu, meningkatkan kualitas data statistik resmi, dan mendukung pengambilan keputusan yang lebih baik di berbagai sektor.*

ABSTRACT. *The Central Statistics Agency (BPS) welcomes the challenge of utilizing big data. One of the BPS publications that can be supported using big data is the inflation figure collected from the consumer price survey. One part of the consumer price survey is the HK-4 Survey, which contains house contract rates. So far, the house contract rates produced by BPS have been underestimated or lower than the actual situation. Improvements to house contract rates are carried out by matching BPS data and web scraping of house rental sites using Propensity Score Matching (PSM). The data used in this study includes DKI Jakarta, Bandung, and Semarang from September to October 2023. This study aims to find the best matching model using PSM to improve official statistics (house contract rates) by combining several propensity score value estimation methods and matching algorithms. Furthermore, the results matching the best model will be used to calculate the corrected house contract rates. The study results show that the best matching model generally uses logistic regression propensity score value estimation, the nearest neighbor matching algorithm with returns and uses a 1:1 ratio. The corrected contract rates are far above the official ones (DKI Jakarta corrected 87.27%, Bandung 316.15%, and Semarang 60.04%). Web Scraping allows it to improve official statistics because it is cost and time-saving, enhances the quality of official statistical data, and supports better decision-making in various sectors.*

1. Pendahuluan

Kehadiran revolusi industri 4.0 mengakibatkan perkembangan teknologi dan informasi menjadi semakin cepat. Salah satu teknologi yang berperan besar dalam revolusi industri 4.0 yaitu *big data* [1]. *Big data* adalah sekumpulan data berukuran besar dan kompleks yang dapat digunakan untuk berbagai analisis dan prediksi. Badan Pusat Statistik (BPS) sebagai Kantor Statistik Resmi menyambut baik dengan kehadiran *big data* ini sebagai opsi untuk meningkatkan pemenuhan kebutuhan data statistik yang berkualitas, lengkap, dan mutakhir. Hal tersebut tertuang dalam salah satu rencana strategis BPS tahun 2020-2024 yaitu: memastikan kemutakhiran data dengan memanfaatkan teknologi informasi dalam pengumpulan data serta penggunaan *big data* untuk mendukung statistik resmi yang dihasilkan [2]. Hal ini juga diperkuat dengan pernyataan Kepala BPS Republik Indonesia dalam kuliah umum yang disampaikan pada tanggal 21 Maret 2023 di Institut Pertanian Bogor (IPB), yang berjudul Kolaborasi Memperkuat Literasi dan Pemanfaatan *Official Statistics*. Pada kuliah umum tersebut Kepala BPS menyampaikan: "Jika dahulu pengumpulan data dilakukan dengan terjun langsung ke lapangan untuk melakukan sensus, survei, atau meminta data kepada

*Penulis Korespondensi.

Email : *twfatimah@apps.ipb.ac.id* (F. Fatimah)
Homepage : http://ejurnal.ung.ac.id/index.php/jjom/index / E-ISSN : 2656-1344

instansi terkait, maka sesuai perkembangan kemajuan teknologi, sekarang dihadapkan pada tantangan untuk memanfaatkan berbagai macam produsen data di luar statistik resmi yaitu dengan memanfaatkan *big data*" [3]. Bahkan negara-negara di dunia yang tergabung dalam *United Nation Global Working Group (GWG) on Big data for Official Statistics* secara khusus melakukan kajian rutin untuk melakukan *sharing* pengalaman pemanfaatan *big data* untuk mendukung statistik resmi [4]. Begitu pula dengan BPS, selama ini telah melakukan beberapa kajian dalam upaya untuk memanfaatkan *big data*. Beberapa kajian yang dilakukan BPS dengan memanfaatkan *big data* yang bisa digunakan untuk statistik resmi antaralain: data citra satelit untuk menghitung angka kemiskinan [5], data dari *marketplace* untuk menghitung statistik harga [6], *mobile positioning data* untuk menghitung jumlah wisatawan manca negara [7]; *web scraping* situs lowongan kerja untuk mendapatkan data jumlah pencari kerja [8], dsb.

Salah satu publikasi yang bisa ditunjang dengan memanfaatkan *big data* yaitu adalah angka inflasi. Angka inflasi dihimpun dari survei harga konsumen yang dikumpulkan di 90 kota inflasi setiap bulannya. Salah satu bagian dari survei harga konsumen adalah Survei HK-4 yang memuat tarif kontrak rumah. Selama ini tarif kontrak rumah yang dihasilkan oleh BPS dinilai *underestimate* atau dibawah keadaan yang sebenarnya, hal ini bisa dilihat berdasarkan data BPS kenaikan tarif kontrak rumah di kota-kota inflasi di Indonesia dalam kurun waktu 5 tahun terakhir mengalami kenaikan yang sangat lambat [9–11].

Penelitian ini mencoba untuk memanfaatkan *big data* sebagai salah satu opsi untuk memperbaiki statistik resmi yang diduga *underestimate*. Florescu *et al.* [12] menyebutkan salah satu peluang pemanfaatan *big data* terkait statistik resmi yaitu kemampuan *big data* untuk memperbaiki statistik resmi. Kemampuan *big data* untuk memperbaiki statistik resmi yaitu untuk mengatasi kelemahan yang melekat pada survei yang sudah dilakukan, seperti kelemahan pada survei HK-4 yaitu kesulitan dalam memperoleh data tarif kontrak rumah *elite* serta keengganan responden dalam menjawab survei yang sifatnya panel. Perbaikan statistik resmi dengan memanfaatkan *big data* dalam penelitian ini menggunakan *web scraping* dari situs sewa rumah. Hasil *web scraping* tersebut akan digunakan untuk melakukan perbaikan terhadap data tarif kontrak rumah resmi yang dinilai *underestimate*. *Web scraping* dari situs sewa rumah memungkinkan untuk mendapatkan data tarif kontrak rumah beserta karakteristik lainnya seperti luas bangunan, jumlah kamar tidur, jumlah kamar mandi, alamat, dsb.

Perbaikan tarif kontrak rumah dilakukan dengan melakukan *matching* antara kelompok perlakuan (data statistik resmi) dan kelompok kontrol (data hasil *web scraping* situs sewa rumah). Metode *matching* yang dapat digunakan diantaranya dengan menggunakan *Propensity Score Matching* (PSM). Tahapan dalam PSM terdiri dari 2 tahapan utama yaitu: melakukan pendugaan nilai *propensity score* dan menentukan algoritma *matching*. Beberapa metode yang dapat digunakan untuk pendugaan nilai *propensity score* dalam PSM diantaranya dengan menggunakan metode klasik (regresi logistik) atau *machine learning* (*random forest, artificial neural network, support vector machine,* dsb). Cannas dan Arpino [13] melakukan perbandingan kinerja beberapa metode dalam PSM baik dengan metode klasik maupun *machine learning;* hasil penelitiannya menunjukkan bahwa *random forest,* regresi logistik, dan *neural network* menghasilkan performa yang bagus dalam menyeimbangkan kovariat diantara kelompok perlakuan dan kontrol dalam PSM. Tahap selanjutnya yaitu menentukan algoritma *matching*. Beberapa penelitian menggunakan algoritma *matching* yang berbeda dalam PSM diantaranya penelitian oleh Chang dan Kim [14] dan Liu *et al.* [15] yang menggunakan algoritma *nearest neighbor matching*, penelitian oleh Austin dan Small [16] yang menggunakan algoritma *optimal matching,* serta penelitian oleh Wood dan Donnell [17] yang menggunakan algoritma *genetic matching*. Algoritma *nearest neighbor matching* merupakan algoritma yang sederhana dan mudah dipahami sedangkan *optimal matching* merupakan algoritma yang berfokus pada meminimalkan jarak global diantara pasangan yang cocok akan tetapi tidak menjamin menghasilkan kovariat seimbang secara keseluruhan. Sedangkan *genetic matching* menggabungkan kovariat-kovariat yang akan di*matching*kan dan mengoptimalkan untuk mencapai keseimbangan kovariat antara kelompok perlakuan dan kontrol. *Genetic matching* ditemukan menghasilkan keseimbangan kovariat yang lebih baik dibandingkan algoritma *matching* lainnya [18]. Pada penelitian ini selain menggunakan ketiga algoritma *matching* tersebut juga menggunakan metode *matching* dengan dan tanpa pengembalian serta menggunakan rasio pencocokan 1:1 hingga 1:5. Penggunaan metode dengan pengembalian lebih bagus digunakan apabila hanya terdapat sedikit data yang tumpang tindih diantara kelompok perlakuan dan kontrol [19]. Penggunaan rasio 1:1 hingga 1:5 dalam penelitian ini dimaksudkan agar diperoleh tarif kontrak yang berasal dari data *webscraping* seberagam mungkin.

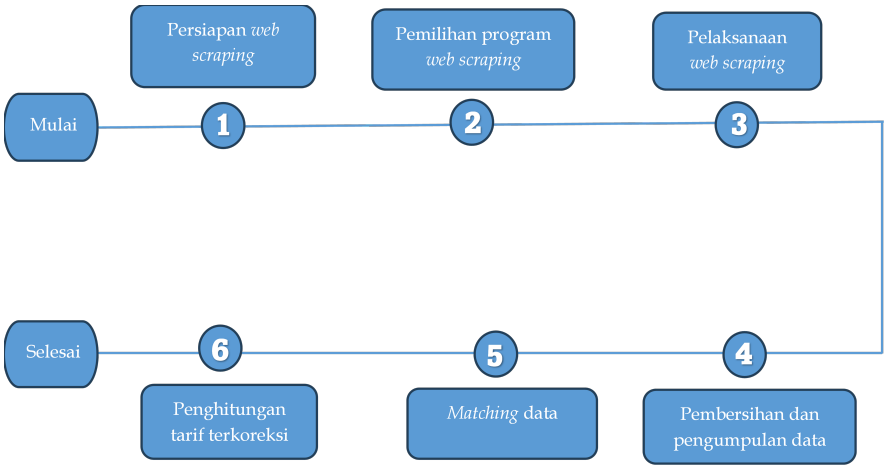
Penelitian ini memperkenalkan pendekatan inovatif untuk memperbaiki statistik resmi khususnya tarif kontrak rumah yang dinilai *underestimate*, dengan menggunakan PSM. Dalam pendekatan ini, menggabungkan metode klasik (regresi logistik) dan *machine learning (random forest dan neural network)* untuk pendugaan nilai *propensity score*. Penelitian ini juga menguji efektivitas tiga *algoritma matching* (*nearest neighbor, optimal,* dan *genetic matching*), serta membandingkan hasil *matching* dengan dan tanpa pengembalian, serta berbagai rasio *matching* untuk mendapatkan data yang lebih representatif. Urgensi penelitian ini terletak pada upaya untuk meningkatkan kualitas data statistik resmi dengan memanfaatkan teknologi *big data,* memberikan solusi praktis untuk mengatasi *underestimate* dalam data, dan mendukung pengambilan keputusan di berbagai sektor.

Berdasarkan uraian tersebut, penelitian bertujuan untuk mencari model *matching* terbaik dengan menggunakan PSM untuk perbaikan statistik resmi yang dinilai *underestimate* dengan mengkombinasikan beberapa metode pendugaan nilai *propensity score* serta menggunakan beberapa algoritma *matching*. Hasil *matching*nya kemudian akan digunakan untuk memperbaiki tarif kontrak resmi.

2. Metode

2.1. Tahapan penelitian

Data yang digunakan dalam penelitian ini berasal dari dua sumber. Data pertama berasal dari hasil survei HK-4 yang dihasilkan oleh BPS yang dikumpulkan dalam rentang waktu September s.d. Oktober 2023 di 3 kota yaitu DKI Jakarta, Bandung, dan Semarang. Data kedua berasal dari hasil *web scraping* situs sewa rumah diantaranya situs: rumah.com, rumah123.com, dan realoka.com yang diposting dalam rentang September s.d. Oktober 2023 di 3 kota yaitu DKI Jakarta, Bandung, dan Semarang. Da-



Gambar 1. *Flowchart* tahapan penelitian

ta hasil survei HK-4 digunakan sebagai kelompok perlakuan dan data hasil *web scraping* sebagai kelompok kontrol. Berdasarkan kedua data tersebut akan dilakukan *matching* data berdasarkan kovariat yang sama diantara kedua data tersebut, dan kovariat yang digunakan dalam penelitian ini disajikan dalam Tabel 1.

Tabel 1. Kovariat penelitian

Kovariat	Penjelasan	Jenis Data
Y	Kelompok (1: perlakuan dan 0: kontrol)	Kategori biner
KT	Jumlah kamar tidur	Numerik
KM	Jumlah kamar mandi	Numerik
LB_1	Luas bangunan 1 (1: lebih dari 250 m^2 dan 0: tidak)	Kategori biner
LB_2	Luas bangunan 2 (1: 101 - 250 m^2 dan 0: tidak)	Kategori biner
LB_3	Luas bangunan 3 (1: 50 - 100 m^2 dan 0: tidak)	Kategori biner
$Garasi$	Garasi (1: ada dan 0: tidak)	Kategori biner

Tahapan penelitian yang digunakan dalam penelitian ini disajikan pada Gambar 1. Tahap pertama adalah tahap persiapan untuk melakukan *web scraping*. Pada tahap ini menentukan situs sewa rumah yang akan dijadikan sebagai sumber data. Situs sewa rumah yang digunakan diantaranya rumah.com, rumah123.com, dan realoka.com. Tahap kedua yaitu menentukan program yang akan digunakan untuk *web scraping*. Program yang digunakan untuk melakukan *web scraping* yaitu menggunakan *python* versi 3.11.5. dengan pustaka *beautifulsoup*. Tahap ketiga yaitu melakukan *web scraping*. *Web scraping* dilakukan untuk memperoleh data-data yang diposting di situs sewa rumah pada rentang september hingga oktober 2023. Tahap keempat yaitu membersihkan dan mengumpulkan data. Pada tahap ini membuang data ganda dan tidak lengkap, serta melakukan pengecekan kebenaran data. Tahap kelima yaitu melakukan *matching* data. Pada tahap ini melakukan *matching* data BPS dan hasil *web scraping* dengan menggunakan PSM. Tahap keenam yaitu menghitung tarif kontrak rumah terkoreksi. Berdasarkan hasil *matching* terbaik dilakukan penghitungan tarif kontrak rumah terkoreksi.

Selanjutnya dipaparkan beberapa konsep penting berkaitan dengan penelitian ini. Konsep-konsep penting yang dipaparkan berikut diperlukan pada pembahasan hasil penelitian.

2.2. *Web Scraping*

Web scraping atau biasa disebut dengan *web extraction* atau *web harvesting* adalah proses pengambilan data/informasi dari internet yang umumnya situs secara otomatis dengan menggunakan *software* [20]. *Web scraping* memungkinkan untuk mengambil data yang yang bersumber dari situs yang sekiranya menarik perhatian untuk dipergunakan kembali dalam berbagai bidang lain [21]. *Web scraping* situs sewa rumah memungkinkan untuk mengambil data yang tersedia di situs sewa rumah seperti: tarif sewa rumah, jumlah kamar tidur, jumlah kamar mandi, alamat, dsb; yang mana dengan data tersebut kemudian akan digunakan untuk perbaikan *official statistics* dalam hal ini tarif kontrak rumah.

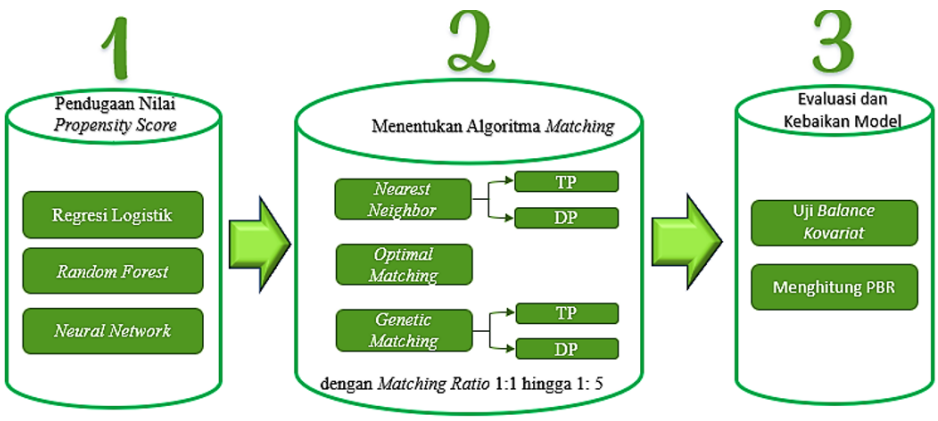
2.3. *Propensity Score Matching (PSM)*

PSM adalah metode yang populer untuk membuat distribusi kovariat yang seimbang diantara kelompok perlakuan dan kontrol. PSM memiliki keunggulan dalam kemampuan mereduksi bias diantara kelompok perlakuan dan kontrol serta kemampuan dalam mereduksi dimensi kovariat hingga menjadi satu nilai yaitu *propensity score* [22, 23]. PSM memiliki 3 tahapan yaitu: 1) melakukan pendugaan nilai *propensity score*, 2) menentukan algoritma *matching,* 3) evaluasi dan kebaikan model. Metode pendugaan nilai *propensity score* dalam PSM yang digunakan dalam penelitian ini diantaranya: regresi logistik, *random forest*, dan *neural network*. Pendugaan nilai *propensity score* dengan menggunakan regresi logistik dilakukan dengan memasukkan persamaan regresi logistik ke dalam persamaan pendugaan *propensity score*. Persamaan pendugaan nilai *propensity score* dengan menggunakan regresi logistik sebagai berikut:

$$PS_i = \frac{e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}. \quad (1)$$

PS_i adalah pendugaan nilai *propensity score* observasi ke-i dengan
i = $1, 2, \ldots, n$;
β_0 : konstanta;
β_i : koefisien regresi;
x_i : kovariat.

Pendugaan nilai *propensity score* dengan menggunakan *random forest* dilakukan dengan cara memasukkan persamaan *random forest* ke dalam persamaan pendugaan nilai *propensity score*. Persamaan pendugaan nilai *propensity score* dengan menggunakan



Gambar 2. Tahapan *matching* data dengan PSM

random forest sebagai berikut:

$$PS_i = \frac{1}{M}\sum_{j=1}^{M} I\left(f_{j(x_i)}\right). \tag{2}$$

PS_i adalah pendugaan nilai *propensity score* observasi ke-i dengan $i = 1, 2, \ldots, n$; I fungsi indikator akan bernilai 1 jika argumen benar dan 0 jika argumen salah; $f_{j(x_i)}$ nilai peluang pohon ke-j; $j = 1, 2, \ldots, M$; M banyaknya pohon. Hasil pendugaan nilai *propensity score* dengan menggunakan *random forest* dilakukan dengan merata-ratakan prediksi dari semua pohon yang terbentuk.

Pendugaan nilai *propensity score* dengan menggunakan *neural network* dilakukan dengan cara memasukkan persamaan fungsi aktivasi non-linear *(sigmoid)* ke dalam persamaan pendugaan nilai *propensity score*. Persamaan pendugaan nilai *propensity score* dengan menggunakan *neural network* sebagai berikut:

$$PS_i = \frac{1}{1+e^{-z_i}}, \tag{3}$$

$$z_i = \sum_i w_i x_i + b. \tag{4}$$

PS_i adalah pendugaan *propensity score* observasi ke-i dengan

i = $1, 2, \ldots, n$;
w_i : bobot kovariat ke-i;
x_i : kovariat ke-i;
b : bias.

Algoritma *matching* yang digunakan dalam penelitian ini meliputi: *nearest neighbor matching, optimal matching,* dan *genetic matching*. Dilakukan dengan dan tanpa pengembalian dan menggunakan rasio 1:1 hingga 1:5. Tahapan *matching* data dengan PSM disajikan dalam Gambar 2.

Algoritma *nearest neighbor matching* melakukan pemadanan berdasarkan nilai pendugaan *propensity score* terdekat, *optimal matching* melakukan pemadanan dengan meminimumkan total jarak *propensity score* diantara unit kelompok perlakuan dan kontrol yang sepadan, sedangkan *genetic matching* melakukan pemadanan dengan memberi bobot untuk masing-masing kovariat dan juga nilai *propensity score*; kemudian bobot-bobot tersebut akan bermutasi sehingga diperoleh bobot-bobot yang mengoptimalkan keseimbangan kovariat [24, 25]. Algoritma *matching* dilakukan dengan dan tanpa pengembalian menggunakan rasio 1:1 hingga 1:5.

Setelah tahap 1 dan 2 dalam PSM dilakukan, tahap selanjutnya adalah mengevaluasi model dengan menggunakan uji *balance* kovariat dan menghitung PBR. Uji *balance* kovariat dalam PSM digunakan untuk memeriksa keseimbangan kovariat antara kelompok perlakuan dan kontrol, jika kovariat antara kelompok perlakuan dan kontrol tidak seimbang, maka dapat menyebabkan penarikan kesimpulan menjadi bias [26]. *Standardized Mean Difference* (SMD) adalah metode statistik yang paling umum digunakan untuk memeriksa keseimbangan distribusi kovariat antara kelompok perlakuan dan kontrol [27]. Persamaan SMD jika kovariat yang diperbandingkan berskala kontinyu menggunakan rumus:

$$|\text{SMD}| = \frac{\overline{x}_t - \overline{x}_k}{\sqrt{(S_t^2 + S_k^2)/2}}. \tag{5}$$

$\overline{x}_t$ dan $\overline{x}_k$ adalah *mean* kovariat dari kelompok perlakuan dan kontrol, S_t^2 dan S_k^2 adalah varians kovariat dari kelompok perlakuan dan kontrol.

Persamaan SMD jika kovariat yang diperbandingkan berskala diskret menggunakan rumus:

$$|\text{SMD}| = \frac{p_p - p_k}{\sqrt{\frac{[p_t(1-p_t)+p_k(1-p_k)]}{2}}}, \tag{6}$$

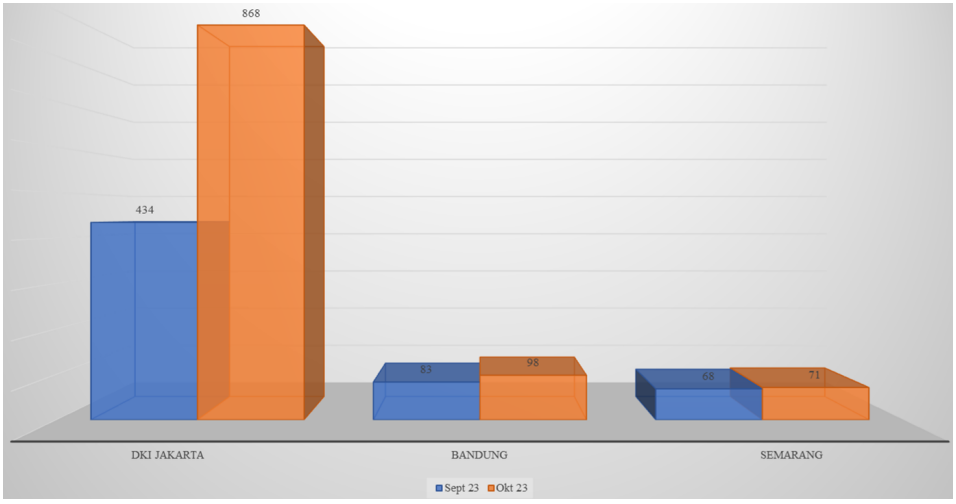
dengan p_t dan p_k adalah proporsi kovariat dari kelompok perlakuan dan kontrol.

Nilai absolut SMD yang lebih kecil dari 0,25 menunjukkan adanya keseimbangan distribusi kovariat antara kelompok perlakuan dan kontrol [28]. Jika ada kovariat yang tidak seimbang, maka kovariat tersebut dikeluarkan dari model dan langkah selanjutnya akan dilakukan penghitungan *Percent Bias Reduction* (PBR). PBR digunakan untuk menghitung berapa persen bias yang tereduksi dari sebelum dilakukan *matching* sampai setelah dilakukan *matching* [29]. Persamaan PBR adalah sebagai berikut:

$$\text{PBR} = \frac{B_b - B_a}{B_b} \; x\; 100\%, \tag{7}$$

$$B = |M(X_t) - M(X_k)|. \tag{8}$$

B adalah *bias*, B_b adalah bias sebelum *matching*, B_a adalah bias sesudah *matching*, M adalah *mean*, X_t adalah kovariat pada kelompok perlakuan, X_k adalah kovariat pada kelompok kontrol. Model dengan kovariat seimbang terbanyak dan dengan nilai PBR tertinggi merupakan model yang terbaik.



Gambar 3. Jumlah amatan hasil *web scraping* situs sewa rumah

2.4. Penghitungan Tarif Kontrak Rumah Terkoreksi

Tarif kontrak rumah adalah besaran tarif sewa tempat tinggal yang dilakukan oleh KRT/ART dengan perjanjian dan batasan waktu tertentu. Tarif kontrak rumah terkoreksi dihitung setelah diperoleh data hasil *matching* yang diperoleh dengan model terbaik. Penghitungan tarif kontrak rumah terkoreksi sama dengan perhitungan tarif kontrak rumah resmi. Penghitungan tarif kontrak rumah resmi dilakukan dengan 2 tahap [30]. Tahap pertama menghitung tarif kontrak rumah menurut jenis kategori rumah dengan rumus sebagai berikut:

$$P_i = \sqrt[n]{x_{i1}.x_{i2}.x_{i3}. \ldots .x_{in}}. \tag{9}$$

P_i adalah tarif kontrak rumah pada kategori i ; $i = 1, 2, \ldots, 16$; x_{ij} adalah tarif kontrak rumah pada kategori i sampel ke -j ; $j = 1, 2, \ldots, n$.

Kategori rumah yang ada dalam perhitungan tarif kontrak rumah resmi terdapat 16 kategori. Pembentukan kategori didasarkan pada 2 komponen yaitu kondisi bangunan (jenis atap, jenis dinding, jenis lantai, luas lantai) dan fasilitas (jenis air, daya listrik, kepemilikan garasi, kepemilikan AC).

Tahap kedua yaitu menghitung tarif kontrak rumah secara agregat. Rumus penghitungan tarif kontrak rumah agregat sebagai berikut:

$$P_a = \prod_{i=1}^{16} (P_i^{b_i}), \tag{10}$$

$$b_i = \frac{s_i}{\sum_{i=1}^{16} s_i}. \tag{11}$$

P_a adalah tarif kontrak rumah agregat, P_i adalah tarif kontrak rumah pada kategori i; $i = 1, 2, \ldots, 16$; b_i adalah bobot pada kategori i, s_i adalah jumlah sampel pada kategori i.

Dikarenakan pada data *web scraping* tidak memiliki data yang lengkap untuk melakukan pengkategorian yang serupa, maka pada penelitian ini begitu dilakukan *matching*, pengkategorian dan pembobotan menggunakan kategori dan bobot yang digunakan pada perhitungan tarif kontrak rumah resmi.

3. Hasil dan Pembahasan

3.1. Data Hasil Web Scraping

Data yang digunakan pada penelitian ini berasal dari data BPS dan hasil *web scraping* situs sewa rumah. Jumlah amatan pada data BPS yaitu masing-masing 26 amatan untuk DKI Jakarta dan Semarang; sedangkan untuk Bandung yaitu sebanyak 35 amatan. Jumlah amatan hasil *web scraping* setelah dikurangi yang ganda dan tidak lengkap disajikan pada Gambar 3.

Jumlah amatan hasil *web scraping* untuk DKI Jakarta paling banyak dibanding Kota Bandung dan Semarang. Jumlah amatan hasil *web scraping* untuk DKI Jakarta pada bulan Oktober 2023 dua kali lipat dibandingkan pada bulan September 2023, sedangkan untuk Kota Bandung dan Semarang jumlah amatan hasil *web scraping* pada bulan September dan Oktober 2023 hanya terdapat sedikit perbedaan. Jumlah amatan hasil *web scraping* jika dirinci menurut bulan, kota dan situs sewa rumah selengkapnya disajikan pada Tabel 2.

Tabel 2. Hasil *web scraping* sesudah dikurangi data yang ganda dan tidak lengkap September – Oktober 2023

Bulan	Kota	Situs Sewa Rumah		
		rumah.com	rumah123.com	realoka.com
September	DKI Jakarta	311	112	11
	Bandung	68	13	2
	Semarang	5	59	4
Oktober	DKI Jakarta	391	427	50
	Bandung	70	15	13
	Semarang	10	52	9

Jumlah amatan hasil *web scraping* di DKI Jakarta dan Bandung pada bulan September 2023 paling banyak pada situs rumah.com dibanding situs lainnya; sedangkan di Semarang jumlah amatan hasil *web scraping* di rumah123.com paling banyak dibanding situs lainnya. Pada bulan Oktober 2023 jumlah amatan hasil *web scraping* di DKI Jakarta dan Semarang pada situs rumah123.com paling banyak dibanding situs lainnya; sedangkan di Bandung jumlah amatan hasil *web scraping* paling banyak pada situs rumah.com dibandingkan situs lainnya.

Tabel 3. Hasil *matching* data di DKI Jakarta pada bulan September 2023

Metode Pend.	Algoritma	Metode	Rasio	Kovariat Seimbang	PBR %
Regresi logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99,8770
Regresi logistik	NN	DP	1:3	$KT, KM, LB_1, LB_2, Garasi$	80,1549
Regresi logistik	Gen	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	86,8929
Regresi logistik	Gen	DP	1:2	$KT, KM, LB_1, LB_2, Garasi$	78,1262
Regresi logistik	Gen	DP	1:3	$KT, KM, LB_1, LB_2, Garasi$	79,4418
Random forest	NN	DP	1:1	$KT, KM, LB_1, LB_2, Garasi$	93,3805
Random forest	NN	DP	1:2	$KT, KM, LB_1, LB_2, Garasi$	82,6560
Random forest	Gen	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	78,3472
Random forest	Gen	DP	1:2	$KT, KM, LB_1, LB_2, Garasi$	84,0832
Neural network	Gen	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	9,7441
Neural network	Gen	DP	1:2	$KT, KM, LB_1, LB_2, Garasi$	27,9636
Neural network	Gen	DP	1:3	$KT, KM, LB_1, LB_2, Garasi$	32,4540

Tabel 4. Hasil *matching* data di DKI Jakarta pada bulan Oktober 2023

Metode Pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
Regresi logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99,3108
Regresi logistik	NN	DP	1:2	$KM, LB_1, LB_2, LB_3, Garasi$	98,2700
Regresi logistik	NN	DP	1:3	$KT, KM, LB_1, LB_3, Garasi$	96,0900
Regresi logistik	Gen	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99,3108
Regresi logistik	Gen	DP	1:2	$KM, LB_1, LB_2, LB_3, Garasi$	99,1702
Random forest	NN	DP	1:1	$KT, KM, LB_1, LB_3, Garasi$	79,0448
Random forest	Gen	DP	1:1	$KT, KM, LB_1, LB_3, Garasi$	80,8090
Random forest	Gen	DP	1:2	$KT, KM, LB_1, LB_3, Garasi$	68,7464
Random forest	Gen	DP	1:3	$KT, KM, LB_1, LB_3, Garasi$	66,1503
Neural network	Gen	DP	1:1	$KT, KM, LB_1, LB_3, Garasi$	97,3039

Tabel 5. Hasil *matching* data di Bandung pada bulan September 2023

Metode Pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
Regresi logistik	Gen	DP	1:1	LB_1, LB_2, LB_3	0,4217
Regresi logistik	Gen	DP	1:2	LB_1, LB_2, LB_3	3,0213
Regresi logistik	Gen	DP	1:3	LB_1, LB_2, LB_3	0,4217
Regresi logistik	Gen	DP	1:4	LB_1, LB_2, LB_3	3,0213
Regresi logistik	Gen	nP	1:5	KM, LB_2, LB_3	0,7660
Random forest	Gen	DP	1:1	LB_1, LB_2, LB_3	9,3548
Random forest	Gen	DP	1:3	LB_1, LB_2, LB_3	6,8736
Random forest	Gen	DP	1:4	LB_1, LB_2, LB_3	6,0261
Random forest	Gen	DP	1:5	KM, LB_2, LB_3	4,9442
Neural network	Gen	DP	1:1	LB_1, LB_2, LB_3	0,3174
Neural network	Gen	DP	1:2	LB_1, LB_2, LB_3	1,3235
Neural network	Gen	DP	1:3	LB_1, LB_2, LB_3	0,2521
Neural network	Gen	DP	1:4	KM, LB_1, LB_3	0,2449
Neural network	Gen	DP	1:5	KM, LB_1, LB_3	0,4316

3.2. Hasil Matching Data di DKI Jakarta

Hasil *matching* data di DKI Jakarta pada bulan September 2023 disajikan pada Tabel 3. Hasilnya menunjukkan terdapat 12 model dengan banyaknya kovariat seimbang yang sama, yaitu sebanyak 5 kovariat. Kovariat dianggap seimbang jika nilai SMD *(Standardized Mean Difference)* di bawah 0,25. Metode *matching* terbaik yaitu model yang selain memiliki kovariat seimbang terbanyak juga memiliki PBR tertinggi. Model *matching* terbaik di DKI Jakarta pada bulan September 2023 yaitu model dengan pendugaan nilai *propensity score* regresi logistik dengan algoritma *nearest neigbor matching* dengan pengembalian menggunakan rasio 1:1.

Hasil *matching data* di DKI Jakarta pada bulan Oktober 2023 disajikan pada Tabel 4. Hasil *matching*nya menunjukkan terdapat 10 model dengan jumlah kovariat seimbang yang sama yaitu sebanyak 5 kovariat. Metode *matching* terbaik yaitu model dengan pendugaan nilai *propensity score* regresi logistik dengan algoritma: *nearest neigbor matching* dengan pengembalian rasio 1:1 dan *genetic matching* dengan pengembalian menggunakan rasio 1:1 dan 1:2.

3.3. Hasil Matching Data di Bandung

Hasil *matching data* di Bandung pada bulan September 2023 disajikan pada Tabel 5. Hasil *matching*nya menunjukkan terdapat 14 model dengan banyaknya kovariat seimbang yang sama yaitu sebanyak 3 kovariat. Jika dilihat secara seksama PBR untuk 14 model tersebut kurang dari 10% sehingga hasil *matching* dari 14 model tersebut tidak bisa digunakan untuk menghitung tarif kon-

Tabel 6. Hasil *matching d*ata di Bandung pada bulan Oktober 2023

Metode Pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
Regresi logistik	Gen	DP	1:1	KT, LB_1, LB_2	72,6696
Random forest	NN	DP	1:1	$KT, LB_1, LB_2, Garasi$	88,2404
Random forest	Gen	DP	1:1	KT, LB_1, LB_2	81,0705
Neural network	Gen	DP	1:1	KT, LB_1, LB_2	5,0238

Tabel 7. Hasil *matching* data di Semarang pada bulan September 2023

Metode Pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
Regresi logistik	NN	DP	1:1	$KM, LB_1, LB_2, Garasi$	96,9114
Regresi logistik	Gen	DP	1:2	KM, LB_1, LB_2, LB_3	29,7223
Regresi logistik	Gen	DP	1:3	KM, LB_1, LB_2, LB_3	7,2030
Regresi logistik	Gen	DP	1:4	KM, LB_1, LB_2, LB_3	8,3900
Regresi logistik	Gen	DP	1:5	KM, LB_1, LB_2, LB_3	7,8138
Random forest	Gen	DP	1:2	KT, KM, LB_1, LB_2	20,4682
Random forest	Gen	DP	1:4	KM, LB_1, LB_2, LB_3	33,8365
Neural network	NN	DP	1:1	$KM, LB_1, LB_2, Garasi$	95,8625
Neural network	Gen	DP	1:3	KM, LB_1, LB_2, LB_3	13,7529
Neural network	Gen	DP	1:4	KM, LB_1, LB_2, LB_3	21,0140
Neural network	Gen	DP	1:5	KM, LB_1, LB_2, LB_3	19,6154

Tabel 8. Hasil *matching* data di Semarang pada bulan Oktober 2023

Metode Pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
Regresi logistik	Gen	DP	1:1	KM, LB_1, LB_2, LB_3	2,5752
Random forest	Gen	DP	1:3	KM, LB_1, LB_2, LB_3	8,6161
Random forest	Gen	DP	1:4	KM, LB_1, LB_2, LB_3	9,2469
Neural network	Gen	DP	1:1	KM, LB_1, LB_2, LB_3	2,3615
Neural network	Gen	DP	1:4	KM, LB_1, LB_2, LB_3	0,3495

trak rumah terkoreksi karena semakin kecil nilai PBR menunjukkan antara data BPS dengan data hasil *web scraping* terdapat ketidakmiripan yang sangat besar sehingga seolah-olah kedua data tersebut bukan berasal dari populasi yang sama sehingga apabila dipaksakan biasnya sangat besar.

Hasil *matching data* di Bandung pada bulan Oktober 2023 disajikan pada Tabel 6. Hasil *matching*nya menunjukkan hanya terdapat 1 model dengan jumlah kovariat seimbang yang terbanyak yaitu model dengan pendugaan nilai *propensity score random forest* dengan algoritma *nearest neigbor matching* dengan pengembalian menggunakan rasio 1:1. Model dengan pendugaan nilai *propensity score random forest* dengan algoritma *nearest neigbor matching* dengan pengembalian menggunakan rasio 1:1 merupakan model terbaik karena disamping memiliki kovariat seimbang yang paling banyak, juga memiliki nilai PBR tertinggi yaitu 88,24%. Selengkapnya bisa dilihat pada Tabel 6.

3.4. Hasil Matching Data di Semarang

Hasil *matching data* di Semarang pada bulan September 2023 disajikan pada Tabel 7. Hasil *matching*nya menunjukkan terdapat 11 model dengan jumlah kovariat seimbang yang sama yaitu sebanyak 4 kovariat. Metode *matching* terbaik yaitu model dengan pendugaan *propensity score* regresi logistik dengan algoritma *nearest neigbor matching* dengan pengembalian menggunakan rasio 1:1.

Hasil *matching data* di Semarang pada bulan Oktober 2023 disajikan pada Tabel 8. Hasil *matching*nya menunjukkan terdapat 5 model dengan banyaknya kovariat seimbang yang sama yaitu sebanyak 4 kovariat. Jika dilihat secara seksama PBR untuk 5 model tersebut kurang dari 10% sehingga hasil *matching* dari 5 model tersebut tidak bisa digunakan untuk menghitung tarif kontrak rumah terkoreksi karena semakin kecil nilai PBR menunjukkan antara data BPS dengan data hasil *web scraping* terdapat ketidakmiripan yang sangat besar sehingga seolah-olah kedua data tersebut bukan berasal dari populasi yang sama sehingga apabila dipaksakan biasnya sangat besar.

3.5. Penghitungan Tarif Kontrak Rumah Terkoreksi

Model *matching* terbaik untuk melakukan *matching* data BPS dengan *web scraping* secara umum menggunakan pendugaan nilai *propensity score* regresi logistik, algoritma *nearest neighbor matching* dengan pengembalian menggunakan rasio 1:1. Pada dasarnya pendugaan nilai *propensity score* baik regresi logistik, *random forest*, maupun *neural network* memiliki kemampuan yang bagus dalam menyeimbangkan kovariat diantara kelompok perlakuan dan kontrol. Hal ini sejalan dengan penelitian yang dilakukan oleh Cannas dan Arpino [13]. Hanya saja pendugaan nilai *propensity score* dengan regresi logistik memiliki keunggulan yaitu menghasilkan nilai PBR yang lebih tinggi dibanding metode pendugaan nilai *propensity score* lainnya. PBR yang tinggi berarti hasil *matching* data menunjukkan kemiripan diantara kelompok perlakuan dan kontrol. Keunggulan metode pendugaan nilai *propensity score* dengan regresi logistik dibanding metode lainnya disajikan pada Tabel 9.

Dalam penelitian ini, algoritma *nearest neighbor matching* dan *genetic matching* menunjukkan kemampuan yang setara da-

Tabel 9. Perbandingan model-model PSM untuk melihat kinerja metode pendugaan nilai *propensity score*

Kota	Bulan	Metode pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
DKI Jakarta	Sept	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.8770
DKI Jakarta	Sept	*Random Forest*	NN	DP	1:1	$KT, KM, LB_1, LB_2, Garasi$	93.3805
DKI Jakarta	Okt	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.3108
DKI Jakarta	Okt	*Random Forest*	NN	DP	1:1	$KT, KM, LB_1, LB_3, Garasi$	79.0448
Semarang	Sept	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, Garasi$	96.9114
Semarang	Sept	*Neural Network*	NN	DP	1:1	$KM, LB_1, LB_2, Garasi$	95.8625

Tabel 10. Perbandingan model-model PSM untuk melihat kinerja algoritma *matching*

Kota	Bulan	Metode pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
DKI Jakarta	Sept	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.8770
DKI Jakarta	Sept	Regresi Logistik	Gen	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	86.8929
DKI Jakarta	Okt	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.3108
DKI Jakarta	Okt	Regresi Logistik	Gen	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.3108
DKI Jakarta	Okt	*Random Forest*	NN	DP	1:1	$KT, KM, LB_1, LB_2, Garasi$	93.3805
DKI Jakarta	Okt	*Random Forest*	NN	DP	1:1	$KT, KM, LB_1, LB_3, Garasi$	79.0448
DKI Jakarta	Okt	*Random Forest*	Gen	DP	1:1	$KT, KM, LB_1, LB_3, Garasi$	80.8090
Bandung	Okt	*Random Forest*	NN	DP	1:1	$KT, LB_1, LB_2, Garasi$	88.2404
Bandung	Okt	*Random Forest*	Gen	DP	1:1	KT, LB_1, LB_2	81.0705

Tabel 11. Perbandingan model-model PSM untuk melihat kinerja rasio *matching*

Kota	Bulan	Metode pend.	Algoritma	Metode	Rasio	Kovariat seimbang	PBR %
DKI Jakarta	Sept	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.8770
DKI Jakarta	Sept	Regresi Logistik	NN	DP	1:3	$KT, KM, LB_1, LB_2, Garasi$	80.1549
DKI Jakarta	Okt	Regresi Logistik	NN	DP	1:1	$KM, LB_1, LB_2, LB_3, Garasi$	99.3108
DKI Jakarta	Okt	Regresi Logistik	NN	DP	1:2	$KM, LB_1, LB_2, LB_3, Garasi$	98.2700
DKI Jakarta	Okt	Regresi Logistik	NN	DP	1:3	$KT, KM, LB_1, LB_3, Garasi$	96.0900

lam menyeimbangkan kovariat antara kelompok perlakuan dan kontrol, seperti yang ditunjukkan pada Tabel 10. Meskipun demikian, temuan ini sedikit berbeda dengan hasil penelitian Diamond dan Sekhon [18], yang menunjukkan bahwa algoritma *genetic matching* cenderung lebih unggul dibandingkan algoritma *matching* lainnya dalam mencapai keseimbangan kovariat. Selain itu, algoritma *nearest neighbor matching* menghasilkan PBR yang lebih tinggi dibandingkan *genetic matching*.

PSM dengan pengembalian ini dimungkinkan dilakukan apabila data pada kelompok perlakuan hanya sedikit tumpang tindih dengan data kelompok kontrol, yang mana apabila hanya sedikit data yang tumpang tindih, maka akan sangat sedikit kovariat yang seimbang yang berakibat hasil *matching* yang dihasilkan tidak optimal. Maka pada penelitian ini penggunakan PSM dengan pengembalian lebih baik hasilnya dibandingkan tanpa pengembalian (Tabel 3-8 menunjukkan model-model yang memiliki kovariat seimbang terbanyak adalah model-model dengan pengembalian). Hal ini sejalan dengan Staffa [19] yang menyarankan penggunaan PSM dengan pengembalian apabila hanya terdapat sedikit data yang tumpang tindih antara kelompok perlakuan dan kontrol.

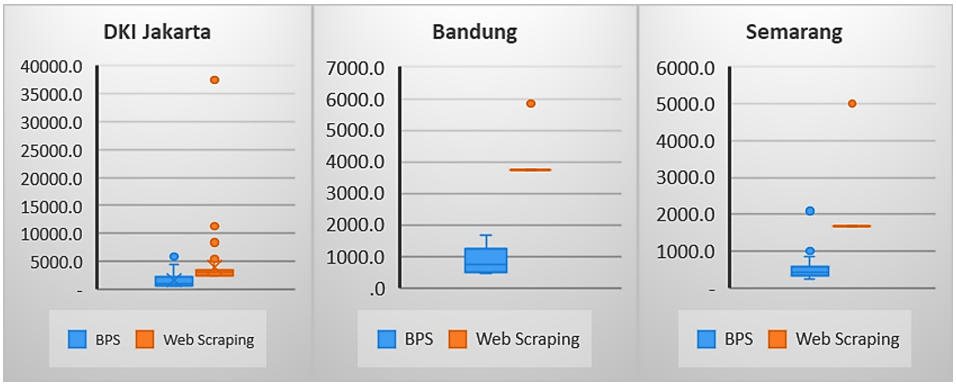
Pada penelitian ini model PSM terbaik diperoleh menggunakan rasio 1:1, hal ini bisa dilihat pada Tabel 11 yang menunjukkan model dengan rasio 1:1 memiliki PBR yang lebih tinggi dibanding model dengan rasio lainnya. Rasio 1:1 berarti misal 1 data BPS dicocokkan dengan 1 data hasil *web scraping*, maka dari 1 data hasil *web scraping* tersebut bisa terpilih kembali untuk *matching* dengan data BPS yang lain. Hal ini disebabkan karena terdapat beberapa data BPS yang berdasarkan kovariatnya memiliki kemiripan karakteristik.

Tabel 12. Perbandingan tarif kontrak resmi dan terkoreksi

Kota	Tarif resmi		Tarif terkoreksi	
	Sept	Okt	Sept	Okt
DKI Jakarta	910.378	910.378	1.553.294	1.871.210
Bandung	417.201	417.410	-	1.737.062
Semarang	584.033	584.033	934.709	-

Setelah diperoleh model terbaik, maka data hasil *matching* tersebut akan dapat digunakan untuk menghitung tarif kontrak rumah terkoreksi. Pada Tabel 12, tarif kontrak terkoreksi terdapat perbedaan yang signifikan dengan tarif kontrak resmi. Tarif kontrak terkoreksi jauh lebih tinggi dibanding tarif kontrak resmi. Tarif kontrak terkoreksi pada bulan September 2023 untuk DKI Jakarta sekitar 1,553 juta rupiah, sedang pada bulan Oktober 2023 tarif kontrak terkoreksi ada 3 tarif dikarenakan terdapat 3 model terbaik. Untuk mendapatkan tarif kontrak terkoreksi pada bulan Oktober, maka ketiga tarif tersebut kemudian dicari rata-rata geometriknya sehingga diperoleh yaitu 1,871 juta rupiah. Tarif kontrak terkoreksi untuk Kota Bandung yaitu 1,737 juta rupiah pada bulan Oktober sedangkan pada bulan September tidak tersedia dikarenakan hasil *matching* datanya tidak bagus. Begitu juga untuk Kota Semarang tarif kontrak terkoreksi pada bulan September 935 ribu rupiah, sedangkan pada bulan Oktober tidak tersedia dikarenakan hasil *matching* yang tidak bagus.

Tarif kontrak resmi pada bulan September dan Oktober menunjukkan tarif yang hampir sama, dengan demikian pada tarif



Gambar 4. *Box plot* tarif kontrak rumah berdasarkan data BPS dan *web scraping* di DKI Jakarta, Bandung, dan Semarang (dalam ribuan)

kontrak terkoreksi seharusnya juga demikian karena jarak antara bulan September dan Oktober sangat dekat. Tarif kontrak terkoreksi untuk DKI Jakarta dilakukan rata-rata geometrik sehingga diperoleh tarif terkoreksi final pada bulan September dan Oktober yaitu 1,705 juta rupiah, sedangkan tarif kontrak terkoreksi final untuk Kota Bandung yaitu 1,737 juta rupiah, dan 935 ribu rupiah untuk Kota Semarang. Tarif kontrak resmi dan terkoreksi final disajikan pada Tabel 13.

Tabel 13. Perbandingan tarif kontrak resmi dan terkoreksi final

Kota	Tarif resmi		Tarif terkoreksi final	
	Sept	Okt	Sept	Okt
DKI Jakarta	910.378	910.378	1.704.857	1.704.857
Bandung	417.201	417.410	1.737.062	1.737.062
Semarang	584.033	584.033	934.709	934.709

Penggunaan data *web scraping* dapat menjadi alternatif yang efektif untuk memperbaiki statistik resmi, khususnya tarif kontrak rumah. *Web scraping* menawarkan berbagai keuntungan, termasuk kemudahan pelaksanaan, penghematan biaya dan waktu dibandingkan dengan survei lapangan langsung [31, 32]. Selain itu, metode ini dapat meningkatkan kualitas data statistik resmi [33], menyediakan solusi praktis untuk mengatasi kekurangan dalam survei [12], serta mendukung pengambilan keputusan yang lebih baik di berbagai sektor. Dengan memanfaatkan data *web scraping*, pihak terkait dapat memperoleh gambaran yang lebih akurat dan terkini mengenai tarif kontrak rumah, yang sangat penting untuk perencanaan dan kebijakan ekonomi.

4. Kesimpulan

Model *matching* terbaik untuk melakukan *matching* data BPS dengan *web scraping* secara umum menggunakan penduga-an nilai *propensity score* regresi logistik, algoritma *nearest neighbor matching* dengan pengembalian menggunakan rasio 1:1. Tarif kontrak terkoreksi secara keseluruhan jauh di atas tarif kontrak resmi. Tarif kontrak terkoreksi bulan September dan Oktober 2023 untuk DKI Jakarta yaitu sebesar 1,705 juta rupiah (terkoreksi 87,27%), Kota Bandung sebesar 1,737 juta rupiah (terkoreksi 316,15%), dan Kota Semarang 935 ribu rupiah (terkoreksi 60,04%). Penggunaan *web scraping* memungkinkan digunakan sebagai salah satu opsi untuk memperbaiki statistik resmi (tarif kontrak rumah), karena kemudahan dalam melakukan *web scraping,* menghemat biaya dan waktu, meningkatkan kualitas data statistik resmi, menawarkan solusi praktis untuk mengatasi kelemahan dalam survei tradisional, dan mendukung pengambilan keputusan yang lebih baik di berbagai sektor.

Kontribusi Penulis. **Fatimah:** Konseptualisasi, metodologi, perangkat lunak, validasi, investigasi, sumber daya, kurasi data, penulisan—penyusunan draf awal, visualisasi. **Hari Wijayanto:** Metodologi, analisis formal, penulisan—tinjauan dan penyuntingan, supervisi. **Farit Mochamad Afendi:** Metodologi, analisis formal, penulisan—tinjauan dan pengeditan, supervisi). Semua penulis telah membaca dan menyetujui versi manuskrip yang diterbitkan.

Ucapan Terima Kasih. Para penulis menyampaikan terima kasih kepada editor dan reviewer atas pembacaan yang cermat, kritik yang mendalam, dan rekomendasi yang praktis untuk meningkatkan kualitas tulisan ini.

Pembiayaan. Penelitian ini tidak menerima pembiayaan eksternal.

Konflik Kepentingan. Para penulis menyatakan tidak ada konflik kepentingan yang terkait dengan artikel ini.

Referensi

[1] C. O. Klingenberg, M. A. V. Borges, and J. A. do V. Antunes, "Industry 4.0: What makes it a revolution? A historical framework to understand the phenomenon," *Technol Soc*, vol. 70, p. 102009, 2022, doi: 10.1016/j.techsoc.2022.102009.

[2] Perka BPS, *Peraturan Kepala BPS No. 36 Tahun 2020 tentang Rencana Strategis Badan Pusat Statistik Tahun 2020-2024*. 2020.

[3] M. Yuwono, "Kolaborasi Memperkuat Literasi dan Pemanfaatan Official Statistics," in *the Public Lecture*, Bogor, Indonesia, Mar. 2023, pp. 1–12.

[4] A. Ashofteh and J. M. Bravo, "Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems," *Stat J IAOS*, vol. 37, no. 3, pp. 771–789, 2021, doi: 10.3233/SJI-210841.

[5] BPS, "Pemodelan Citra Malam Untuk Estimasi Kemiskinan Desa," Jakarta, Indonesia, Aug. 2022.

[6] S. Pramana and S. Mariyah, "Big data implementation for price statistics in Indonesia: Past, current, and future developments," *Stat J IAOS*, vol. 37, no. 1, pp. 415–427, 2021, doi: 10.3233/SJI-200740.

[7] T. K. Lestari, S. Esko, S. E. Sarpono, and R. Rufiadi, "Indonesia's Experience of using Signaling Mobile Positioning Data for Official Tourism Statistics," in *15th world forum on tourism statistics*, Cusco, Peru, 2018. Accessed: Jul. 26, 2024. [Online]. Available: http://www. 15th-tourism-stats-forum.com/papers.html.

[8] BPS, "Kajian Big Data sebagai Pelengkap Data dan Informasi Statistik Sosial," Jakarta, Indonesia, 2020.

[9] Badan Pusat Statistik (BPS), "Harga Konsumen Beberapa Barang Dan Jasa Kelompok Perumahan 82 Kota Di Indonesia 2017," Jakarta, Indonesia, Mar. 2018.

[10] Badan Pusat Statistik (BPS), "Harga Konsumen Beberapa Barang Dan Jasa Kelompok Perumahan Di 82 Kota Di Indonesia 2019," Jakarta, Indonesia, Mar. 2020.

[11] Badan Pusat Statistik (BPS), "Publikasi Harga Konsumen Beberapa Barang dan Jasa Kelompok Perumahan, Air, Listrik, dan Bahan Bakar Rumah Tangga 90 Kota di Indonesia 2021," Jakarta, Indonesia, Mar. 2022.

[12] D. Florescu, M. Karlberg, F. Reis, P. R. Del Castillo, M. Skaliotis, and A. Wirthmann, "Will 'big data' transform official statistics," in *European Conference on the Quality of Official Statistics. Vienna, Austria*, 2014, pp. 2–5.

[13] M. Cannas and B. Arpino, "A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting," *Biometrical Journal*, vol. 61, no. 4, pp. 1049–1072, Jul. 2019, doi: 10.1002/bimj.201800132.

[14] T. W. Chang and Y. Kim, "Performance analysis of promotion programs of the smart factory using propensity score matching," *Procedia Comput Sci*, vol. 232, pp. 1909–1917, 2024, doi: 10.1016/j.procs.2024.02.013.

[15] W. D. Liu *et al.*, "Effect of early dexamethasone on outcomes of COVID-19: A quasi-experimental study using propensity score matching," *Journal of Microbiology, Immunology and Infection*, vol. 57, no. 3, pp. 414–425, 2024, doi: 10.1016/j.jmii.2024.02.002.

[16] P. C. Austin and D. S. Small, "The use of bootstrapping when using propensity-score matching without replacement: a simulation study," *Stat Med*, vol. 33, no. 24, pp. 4306–4319, 2014, doi: 10.1002/sim.6276.

[17] J. Wood and E. T. Donnell, "Safety evaluation of continuous green T intersections: A propensity scores-genetic matching-potential outcomes approach," *Accid Anal Prev*, vol. 93, pp. 1–13, 2016, doi: 10.1016/j.aap.2016.04.015.

[18] A. Diamond and J. S. Sekhon, "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, vol. 95, no. 3, pp. 932–945, 2013, doi: 10.1162/REST_a_00318.

[19] S. J. Staffa and D. Zurakowski, "Five steps to successfully implement and evaluate propensity score matching in clinical research studies," *Anesth Analg*, vol. 127, no. 4, 2018, doi: 10.1213/ANE.0000000000002787.

[20] B. Zhao, "Web Scraping," in *Encyclopedia of Big Data*, L. A. Schintler and C. L. McNeely, Eds., Cham: Springer International Publishing, 2017, pp. 1–3. doi: 10.1007/978-3-319-32001-4_483-1.

[21] B. Ramsey, M. Turland, and O. Merida, *Web Scraping with PHP, 2nd Edition: A Php[architect] Guide*, 2nd ed. Canada: PHP [Architect], 2019. [Online]. Available: https://books.google.co.id/books?id=OvZryAEACAAJ.

[22] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behav Res*, vol. 46, no. 3, pp. 399–424, 2011, doi: 10.1080/00273171.2011.568786.

[23] S. Guo and M. W. Fraser, *Propensity Score Analysis: Statistical Methods and Applications*, 2nd ed. United States of America: SAGE publications, 2014.

[24] H. Yasunaga, "Introduction to applied statistics—chapter 1 propensity score analysis," *Annals of Clinical Epidemiology*, vol. 2, no. 2, pp. 33–37, 2020, doi: 10.37737/ace.2.2_33.

[25] Q.-Y. Zhao, J.-C. Luo, Y. Su, Y.-J. Zhang, G.-W. Tu, and Z. Luo, "Propensity score matching with R: Conventional methods and new features," *Ann Transl Med*, vol. 9, no. 9, 2021, doi: 10.21037/atm-20-3998.

[26] H. Harris and S. J. Horst, "A brief guide to decisions at each step of the propensity score matching process," *Practical Assessment, Research, and Evaluation*, vol. 21, no. 1, p. 4, 2016, doi: 10.7275/yq7r-4820.

[27] Z. Zhang, H. Kim, G. Lonjon, and Y. Zhu, "Balance diagnostics after propensity score matching," *Ann Transl Med*, vol. 7, p. 16, Jan. 2019, doi: 10.21037/atm.2018.12.10.

[28] D. Bottigliengo, G. Lorenzoni, H. Ocagli, M. Martinato, P. Berchialla, and D. Gregori, "Propensity score analysis with partially observed baseline covariates: A practical comparison of methods for handling missing data," *Int J Environ Res Public Health*, vol. 18, no. 13, p. 6694, 2021, doi: 10.3390/ijerph18136694.

[29] Y. Liu, B. Zumbo, P. Gustafson, Y. Huang, E. Kroc, and A. Wu, "Investigating causal DIF via propensity score methods," *Practical Assessment, Research & Evaluation*, vol. 21, pp. 1–24, Dec. 2016, doi: 10.7275/ewqz-n963.

[30] BPS, "Pedoman dan Pencacahan Survei Tarif Sewa/Kontrak Rumah, Upah Pembantu Rumah Tangga, Upah Baby Sitter, dan Uang Sekolah (STRPBS) 2020," BPS, Jakarta, Indonesia, Oct. 2019.

[31] J. Li, L. Xu, L. Tang, S. Wang, and L. Li, "Big data in tourism research: A literature review," *Tour Manag*, vol. 68, pp. 301–323, 2018, doi: 10.1016/j.tourman.2018.03.009.

[32] J. Irek, "Web scraping for food price research," *British Food Journal*, vol. 121, pp. 3350–3361, Nov. 2019, doi: 10.1108/BFJ-02-2019-0081.

[33] E. L. Groshen, "The future of official statistics," *Harv Data Sci Rev*, vol. 3, no. 4, 2021, doi: 10.1162/99608f92.591917c6.