

PENGEMBANGAN MODEL PROPENSITY SCORE MATCHING PADA PEMANFAATAN DATA HASIL WEB SCRAPING UNTUK PERBAIKAN STATISTIK RESMI: STUDI KASUS TARIF KONTRAK RUMAH

FATIMAH



**PROGRAM STUDI STATISTIKA DAN SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024**



PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Pengembangan Model *Propensity Score Matching* pada Pemanfaatan Data Hasil *Web Scraping* untuk Perbaikan Statistik Resmi: Studi Kasus Tarif Kontrak Rumah” merupakan hasil karya saya sendiri yang disusun dengan arahan dari dosen pembimbing, dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Setiap informasi yang diambil atau dikutip dari karya penulis lain, baik yang diterbitkan maupun tidak, telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Desember 2024

Fatimah
G1501212079



RINGKASAN

FATIMAH. Pengembangan Model *Propensity Score Matching* pada Pemanfaatan Data Hasil *Web Scraping* untuk Perbaikan Statistik Resmi: Studi Kasus Tarif Kontrak Rumah. Dibimbing oleh HARI WIJAYANTO dan FARIT MOCHAMAD AFENDI.

Revolusi Industri 4.0 telah mempercepat perkembangan teknologi dan informasi, dengan *big data* menjadi salah satu teknologi kunci dalam era ini. *Big data* mengacu pada kumpulan data yang sangat besar dan kompleks. Badan Pusat Statistik (BPS) menyambut kehadiran *big data* sebagai upaya untuk meningkatkan kualitas, kelengkapan, dan kemitakhiran data statistik. Hal ini tertuang dalam rencana strategis BPS 2020-2024 yang menyoroti penggunaan teknologi informasi dan *big data* untuk mendukung statistik resmi.

Salah satu publikasi BPS yang bisa ditunjang dengan memanfaatkan *big data* yaitu 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 di bawah 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.

Penelitian ini memanfaatkan *big data* untuk memperbaiki statistik resmi yang diduga *underestimate*. Perbaikan statistik resmi (tarif kontrak rumah) dalam penelitian ini dilakukan dengan menggunakan hasil *web scraping* dari situs sewa rumah. Proses perbaikan tarif kontrak rumah dilakukan dengan *matching* data statistik resmi (kelompok perlakuan) dengan data hasil *web scraping* (kelompok kontrol) dengan menggunakan *Propensity Score Matching* (PSM). PSM terdiri dari dua tahap utama: pendugaan nilai *propensity score* dan penentuan algoritma *matching*. Beberapa metode yang dapat digunakan untuk pendugaan nilai *propensity score* antara lain regresi logistik, *random forest*, dan *neural network*. Berdasarkan penelitian ketiga metode tersebut menghasilkan performa yang bagus dalam menyeimbangkan kovariat antara kelompok perlakuan dan kontrol. Tahap selanjutnya yaitu menentukan algoritma *matching*. Penelitian ini menggunakan beberapa algoritma *matching* yaitu: *nearest neighbor matching*, *optimal matching*, dan *genetic matching*. Menggunakan metode *matching* dengan dan tanpa pengembalian serta menggunakan rasio *matching* 1:1 hingga 1:5.

Tujuan penelitian ini adalah mencari model *matching* terbaik dengan menggunakan *Propensity Score Matching* (PSM) untuk perbaikan statistik resmi yang dinilai *underestimate*, dengan mengkombinasikan beberapa metode pendugaan nilai *propensity score* serta menggunakan beberapa algoritma *matching*. Selain itu, penelitian ini juga bertujuan menghitung tarif kontrak terkoreksi berdasarkan hasil *matching* terbaik.

Penerapan *Propensity Score Matching* (PSM) dalam melakukan *matching* data antara data BPS dengan data hasil *web scraping* situs sewa rumah memberikan hasil yang cukup memuaskan. Hal ini terlihat dari banyaknya kovariat seimbang dan nilai PBR yang tinggi. Penelitian ini menunjukkan bahwa pendugaan nilai *propensity score* menggunakan regresi logistik, *random forest*, dan *neural network*

efektif dalam menyeimbangkan kovariat, akan tetapi regresi logistik menunjukkan performa terbaik dalam hal PBR (*Percent Bias Reduction*). Selain itu algoritma *nearest neighbor* dan *genetic matching* memiliki kemampuan yang setara, namun *nearest neighbor matching* menghasilkan PBR yang lebih baik. Penggunaan PSM dengan pengembalian menghasilkan kovariat seimbang lebih banyak dibandingkan tanpa pengembalian. Model dengan rasio 1:1 pada penelitian ini menghasilkan PBR yang lebih tinggi dibandingkan dengan rasio lainnya.

Tarif kontrak yang telah terkoreksi secara keseluruhan jauh lebih tinggi dibandingkan tarif kontrak resmi. Pada bulan September dan Oktober 2023, tarif kontrak terkoreksi untuk DKI Jakarta yaitu sebesar Rp1,705 juta per bulan (terkoreksi 87,27%), Bandung sebesar Rp1,737 juta per bulan (terkoreksi 316,15%), dan Semarang Rp935 ribu per bulan (terkoreksi 60,04%). Tarif kontrak terkoreksi di Bandung lebih tinggi dibandingkan DKI Jakarta, berbeda dengan tarif resmi di mana DKI Jakarta tertinggi. Perbedaan ini kemungkinan disebabkan oleh kecenderungan rumah yang diunggah pada situs sewa rumah di Bandung didominasi oleh rumah-rumah mewah, sementara rumah biasa relatif sedikit. Selain itu, hasil *matching* di Bandung menunjukkan tingkat kemiripan 88,24%, berbeda dengan DKI Jakarta dan Semarang yang mendekati 100%. Hal ini berdampak pada tarif kontrak yang lebih tinggi di Bandung, sehingga penggunaan tarif kontrak terkoreksi di Bandung perlu kehati-hatian karena adanya ketidaklaziman pada hasil *matching*.

Kata kunci: *big data*, *propensity score matching*, statistik resmi, tarif kontrak rumah, *web scraping*

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



SUMMARY

FATIMAH. Development of a Propensity Score Matching Model in the Utilization of Data from Web Scraping for Improving Official Statistics: A Case Study on Housing Contract Rates. Supervised by HARI WIJAYANTO and FARIT MOCHAMAD AFENDI.

The Industrial Revolution 4.0 has accelerated the development of technology and information, with big data becoming one of the key technologies in this era. Big data refers to a vast and complex collection of data. The Central Statistics Agency (BPS) welcomes the presence of big data as an effort to improve the quality, completeness, and timeliness of statistical data, as stated in the BPS 2020-2024 strategic plan, which highlights the use of information technology and big data to support official statistics.

One of BPS's publications that big data can support is the inflation rate. The inflation rate is collected from a monthly consumer price survey conducted in 90 inflation cities. One part of the consumer price survey is the HK-4 Survey, which contains housing contract rates. So far, the housing contract rates produced by BPS have been considered underestimated or below the actual situation. Based on BPS data, the increase in housing contract rates in cities and inflation in Indonesia over the last five years has been very slow.

This study uses big data to improve official statistics suspected of being underestimated. This study's official statistics (house contract rates) were improved using web scraping results from house rental sites. The process of enhancing house contract rates was carried out by matching official statistical data (treatment group) with web scraping data (control group) using Propensity Score Matching (PSM). PSM consists of two main stages: estimating the propensity score value and determining the matching algorithm. Several methods can be used to estimate the propensity score value, including logistic regression, random forest, and neural networks. Research shows that the three methods balance covariates between the treatment and control groups. The next stage is to determine the matching algorithm. This study uses several matching algorithms: nearest neighbor matching, optimal matching, and genetic matching. Matching methods with and without replacement and a matching ratio of 1: 1 to 1: 5.

This study aims to find the best matching model using Propensity Score Matching (PSM) to improve official statistics that are considered underestimated by combining several propensity score estimation methods and using several matching algorithms. In addition, this study also aims to calculate the corrected contract rate based on the best matching results.

The application of Propensity Score Matching (PSM) in matching data between BPS data and data obtained from web scraping rental housing websites has yielded entirely satisfactory results, as can be seen from many balanced covariates and high PBR values. This study demonstrates that logistic regression, random forest, and neural networks effectively balance covariates when estimating propensity scores. However, logistic regression shows the best performance in terms of Percent Bias Reduction (PBR). In addition, the nearest neighbor and genetic matching algorithms have equivalent capabilities, but nearest neighbor matching produces better PBR. The use of PSM with returns produces more



balanced covariates than without returns. This study's model with a 1:1 ratio creates a higher PBR than other ratios.

The corrected rental rates are overall significantly higher than the official rental rates. In September and October 2023, the corrected rental rate for DKI Jakarta was Rp1.705 million per month (an increase of 87.27%), for Bandung Rp1.737 million per month (an increase of 316.15%), and for Semarang Rp935 thousand per month (an increase of 60.04%). The corrected rental rates in Bandung were higher than in Jakarta, contrasting with the official rates where Jakarta had the highest values. This difference is likely caused by the tendency for luxury homes to dominate the rental property listings in Bandung, while regular homes are relatively few. Additionally, the matching results for Bandung show a similarity rate of 88.24%, which differs from DKI Jakarta and Semarang, where the rates are close to 100%. As a result, rental rates in Bandung are higher; using corrected rental rates in the area requires caution due to the irregularities in the matching results.

Keywords: big data, housing contract rates, propensity score matching, official statistics, web scraping



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2024
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

**PENGEMBANGAN MODEL PROPENSITY SCORE MATCHING
PADA PEMANFAATAN DATA HASIL WEB SCRAPING UNTUK
PERBAIKAN STATISTIK RESMI: STUDI KASUS TARIF
KONTRAK RUMAH**

FATIMAH

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI STATISTIKA DAN SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024**



@Hak cipta milik IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Penguji Luar Komisi Pembimbing pada Ujian Tesis:
Dr. Bagus Sartono S.Si, M.Si.

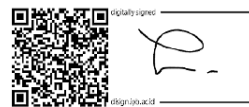


Judul Tesis : Pengembangan Model *Propensity Score Matching* pada
Pemanfaatan Data Hasil *Web Scraping* untuk Perbaikan Statistik
Resmi: Studi Kasus Tarif Kontrak Rumah

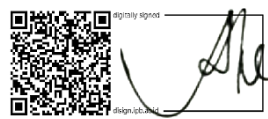
Nama : Fatimah
NIM : G1501212079

Disetujui oleh

Pembimbing 1:
Prof. Dr. Ir. Hari Wijayanto M.Si.



Pembimbing 2:
Dr. Farit Mochamad Afendi S.Si., M.Si.



Diketahui oleh

Ketua Program Studi:
Dr. Agus Mohamad Soleh, S.Si, M.T.
NIP 197503151999031004



Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam:
Dr. Berry Juliandi S.Si., M.Si.
NIP 19780723200701101



Tanggal Ujian:
20 Desember 2024

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa ta'ala* atas segala karunia-Nya sehingga penelitian berjudul “Pengembangan Model *Propensity Score Matching* pada Pemanfaatan Data Hasil *Web Scraping* untuk Perbaikan Statistik Resmi: Studi Kasus Tarif Kontrak Rumah” ini dapat diselesaikan.

Penulis menyadari bahwa penyusunan penelitian ini dapat diselesaikan dengan baik berkat bantuan, dukungan, dan arahan dari berbagai pihak. Oleh karena itu, dengan penuh rasa hormat, penulis menyampaikan terima kasih kepada:

1. Bapak Prof. Dr. Ir. Hari Wijayanto M.Si. dan Bapak Dr. Farit Mochamad Afendi S.Si., M.Si. selaku pembimbing yang senantiasa sabar dan bersedia meluangkan waktunya untuk membimbing, memberikan saran, dan memberikan masukan kepada penulis sehingga penelitian ini dapat diselesaikan dengan baik.
2. Pimpinan, pengajar, dan pegawai Departemen Statistika yang telah membantu terkait penelitian ini.
3. Badan Pusat Statistik yang telah berkenan memberikan datanya untuk kepentingan penelitian ini.
4. Sahabat, teman, dan rekan-rekan mahasiswa pascasarjana Program Studi Statistika dan Sains Data.
5. Teristimewa untuk kedua orang tua dan keluarga kecil (Ayah, Ifa, Afif, Rey) yang saya cintai dan selalu memberikan doa, semangat, dan dukungan dalam menyelesaikan penelitian ini.

Semoga karya ilmiah ini dapat bermanfaat bagi kemajuan ilmu pengetahuan. Penulis juga menyadari bahwa masih terdapat kekurangan dalam penyusunan penelitian ini. Akhir kata, semoga Allah *subhanahu wa ta'ala* senantiasa memberikan rahmat dan hidayah-Nya kepada kita semua. *Aamiin*.

Bogor, Desember 2024

Fatimah



- 

DAFTAR ISI

DAFTAR TABEL	xvii
DAFTAR GAMBAR	xvii
DAFTAR LAMPIRAN	xviii
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan	3
II TINJAUAN PUSTAKA	4
2.1 <i>Web Scraping</i>	4
2.2 <i>Propensity Score Matching (PSM)</i>	4
2.3 Perhitungan Tarif Kontrak Rumah Terkoreksi	11
III METODE	13
3.1 Data	13
3.2 Tahapan Analisis Data	14
3.3 Tahapan Melakukan <i>Web Scraping</i>	15
3.4 Tahapan Melakukan <i>Matching Data</i>	19
3.5 Tahapan Melakukan Perhitungan Tarif Kontrak Terkoreksi	21
IV HASIL DAN PEMBAHASAN	22
4.1 Karakteristik Data BPS dan Hasil <i>Web Scraping</i>	22
4.2 Hasil <i>Matching Data</i> di DKI Jakarta	28
4.3 Hasil <i>Matching Data</i> di Bandung	30
4.4 Hasil <i>Matching Data</i> di Semarang	31
4.5 Perhitungan Tarif Kontrak Rumah Terkoreksi	32
V SIMPULAN DAN SARAN	38
5.1 Simpulan	38
5.2 Saran	38
DAFTAR PUSTAKA	40
LAMPIRAN	43
RIWAYAT HIDUP	57



DAFTAR TABEL

1	Algoritma <i>optimal matching</i>	9
2	Kovariat penelitian	13
3	Hasil <i>web scraping</i> sesudah dikurangi data yang ganda dan tidak lengkap menurut situs sewa rumah dan bulan	23
4	Hasil <i>matching</i> data di DKI Jakarta pada bulan September 2023	29
5	Hasil <i>matching</i> data di DKI Jakarta pada bulan Oktober 2023	29
6	Hasil <i>matching</i> data di Bandung pada bulan September 2023	30
7	Hasil <i>matching</i> data di Bandung pada bulan Oktober 2023	31
8	Hasil <i>matching</i> data di Semarang pada bulan September 2023	31
9	Hasil <i>matching</i> data di Semarang pada bulan Oktober 2023	32
10	Perbandingan model-model PSM untuk melihat kinerja metode pendugaan nilai <i>propensity score</i>	32
11	Perbandingan model-model PSM untuk melihat kinerja <i>nearest neighbor</i> , <i>optimal</i> , dan <i>genetic matching</i>	33
12	Perbandingan model-model PSM untuk melihat kinerja metode dengan dan tanpa pengembalian	34
13	Perbandingan model-model PSM untuk melihat kinerja rasio <i>matching</i>	34
14	Perbandingan tarif kontrak resmi dan terkoreksi per bulan (dalam Rp)	35
15	Perbandingan tarif kontrak resmi dan terkoreksi final per bulan (dalam Rp)	35

DAFTAR GAMBAR

1	Proses <i>web scraping</i>	4
2	Tahapan <i>matching</i> data dengan PSM	5
3	Algoritma <i>random forest</i>	6
4	Algoritma <i>neural network</i>	7
5	Algoritma <i>nearest neighbor matching</i>	8
6	Algoritma <i>genetic matching</i>	10
7	Tahapan analisis data	14
8	Tampilan salah satu rumah yang disewakan di DKI Jakarta pada situs rumah123.com	15
9	Elemen yang diambil pada rumah sewa yang diunggah pada situs rumah123.com	16
10	Tampilan beberapa rumah yang disewakan di DKI Jakarta pada situs realoka.com	17
11	Tampilan salah satu rumah yang disewakan di DKI Jakarta pada situs realoka.com	18
12	Elemen yang diambil pada daftar rumah sewa yang diunggah pada situs realoka.com	19
13	Ilustrasi pendugaan nilai <i>propensity score</i>	20
14	Ilustrasi <i>matching</i> data menggunakan <i>nearest neighbor matching</i> tanpa pengembalian dengan rasio 1:1	21
15	Jumlah amatan hasil <i>web scraping</i> menurut kota dan bulan	22
16	Perbandingan distribusi tarif kontrak rumah per bulan di DKI Jakarta, Bandung, dan Semarang menurut sumber data dan bulan (Rp Jutaan)	24

17	Perbandingan jumlah kamar tidur rumah yang dikontrakkan di DKI Jakarta, Bandung, dan Semarang menurut sumber data dan bulan	25
18	Perbandingan jumlah kamar mandi rumah yang dikontrakkan di DKI Jakarta, Bandung, dan Semarang menurut sumber data dan bulan	26
19	Perbandingan luas bangunan rumah yang dikontrakkan di DKI Jakarta, Bandung, dan Semarang menurut sumber data dan bulan	27
20	Perbandingan kepemilikan garasi rumah yang dikontrakkan di DKI Jakarta, Bandung, dan Semarang menurut sumber data dan bulan	28
21	Perbandingan tarif kontrak resmi dan terkoreksi final per bulan di DKI Jakarta, Bandung, dan Semarang (dalam ribuan)	36

DAFTAR LAMPIRAN

1	Pengecekan kebenaran data yang ada pada situs sewa rumah	43
2	Hasil <i>matching</i> data di DKI Jakarta pada bulan September 2023	45
3	Hasil <i>matching</i> data di DKI Jakarta pada bulan Oktober 2023	47
4	Hasil <i>matching</i> data di Bandung pada bulan September 2023	49
5	Hasil <i>matching</i> data di Bandung pada bulan Oktober 2023	51
6	Hasil <i>matching</i> data di Semarang pada bulan September 2023	53
7	Hasil <i>matching</i> data di Semarang pada bulan Oktober 2023	55