OPTIMASI METODE *RANDOM FOREST* MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK MEMPREDIKSI HARGA RUMAH

THESIS

Oleh:
EMHA AHDAN FAHMI ELMUNA
NIM. 210605220005

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

OPTIMASI METODE *RANDOM FOREST* MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK MEMPREDIKSI HARGA RUMAH

THESIS

Diajukan Kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M. Kom)

Oleh:
EMHA AHDAN FAHMI ELMUNA
NIM. 210605220005

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

OPTIMASI METODE *RANDOM FOREST* MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK MEMPREDIKSI HARGA RUMAH

THESIS

Diajukan Kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M. Kom)

Oleh:
EMHA AHDAN FAHMI ELMUNA
NIM. 210605220005

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

OPTIMASI METODE *RANDOM FOREST* MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK MEMPREDIKSI HARGA RUMAH

THESIS

Oleh:
EMHA AHDAN FAHMI ELMUNA
NIM. 210605220005

Telah diperiksa dan disetujui untuk di uji:
Tanggal 12 Mei 2023

Pembimbing I,	Pembimbing II,
Dr. Totok Chamidy, M. Kom	Dr. Fresy Nugroho, M. T
NIP. 19691222 200604 1 001	NIP. 19710722 201101 1 001

Mengetahui,
Kepala Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

OPTIMASI METODE *RANDOM FOREST* MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK MEMPREDIKSI HARGA RUMAH

THESIS

Oleh:
EMHA AHDAN FAHMI ELMUNA
NIM. 210605220005

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal 12 Mei 2023

Susunan Dewan Penguji			Tanda Tangan
Penguji Utama	:	Dr. Yunifa Miftachul Arif, M.T NIP 19830616 201101 1 004	( )
Ketua Penguji	:	Dr. M. Faisal, M.T NIP. 19740510 200501 1 007	( )
Sekretaris Penguji	:	Dr. Fresy Nugroho, M. T NIP. 19710722 201101 1 001	( )
Anggota Penguji	:	Dr. Totok Chamidy, M. Kom NIP. 19691222 200604 1 001	( )

Mengetahui dan Mengesahkan,
Kepala Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. [illegible]ahyo Crysdian
NIP. 19740424 200901 1 008

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama	: Emha Ahdan Fahmi Elmuna
NIM	: 210605220005
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Thesis yang saya tulis ini benar-banar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka. Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 08 Mei 2023
Yang membuat pernyataan,

Emha Ahdan Fahmi Elmuna
NIM. 210605220005

MOTTO

"Kamu bukanlah tidak mampu,
Hanya saja kamu terlalu menakuti dirimu sendiri,
Dengan kemungkinan yang belum tentu terjadi"

PERSEMBAHAN

Dengan mengucapkan syukur Alhamdulillahi rabbil alamin, Thesis ini saya persembahkan untuk :

1. Seluruh Keluarga tercinta yang selalu memberikan doa dan semangat.
2. Seluruh Civitas Akademika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah memberikan kesempatan untuk menambah ilmu teknologi dan agama.
3. Seluruh rekan-rekan mahasiswa Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang semua angkatan atas kerjasamanya selama ini.
4. Bapak, ibu, saudara dan rekan-rekan sekalian yang tidak bisa saya sebutkan satu persatu dalam mendukung Thesis ini hingga bisa diselesaikan.

KATA PENGANTAR

Assalamu'alaikum Wr. Wb

Syukur *Alhamdulillah* penulis haturkan kehadirat Allah SWT yang telah melimpahkan Rahmat dan Hidayah-Nya, sehingga penulis dapat menyelesaikan studi di Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang sekaligus menyelesaikan Thesis ini dengan baik.

Selanjutnya penulis haturkan ucapan terima kasih seiring do'a dan harapan *jazakumullah ahsanal jaza'* kepada semua pihak yang telah membantu terselesaikannya Thesis ini. Ucapan terima kasih ini penulis sampaikan kepada:

1. Keluarga tercinta yang senantiasa memberikan do'a dan semangat
2. Bapak Dr. Totok Chamidy, M. Kom dan Bapak Dr. Fresy Nugroho, M. T, selaku dosen pembimbing Thesis, dan bapak Dr. Cahyo Crysdian selaku ketua jurusan magister informatika, yang telah banyak memberikan pengarahan dan pengalaman yang berharga.
3. Segenap civitas akademika Program Studi Magister Informatika, terutama seluruh Bapak / Ibu dosen, terima kasih atas segenap ilmu dan bimbingannya.
4. Semua rekan-rekan seperjuangan yang ikut mendukung dan membantu.

Penulis menyadari bahwa dalam penyusunan Thesis ini masih terdapat kekurangan dan penulis berharap semoga Thesis ini bisa memberikan manfaat kepada para pembaca khususnya bagi penulis secara pribadi. *Amiinn Yaa Rabbal Alamin.*

Wasalamu'alaikum Wr. Wb

Malang, 08 Mei 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR RUMUS

ABSTRAK

Ahdan, Emha. 2023. **Optimasi Metode *Random Forest* Menggunakan *Principal Component Analysis* untuk Memprediksi Harga Rumah**. Thesis. Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Totok Chamidy, M. Kom (II) Dr. Fresy Nugroho, M. T

Kata Kunci: Prediksi harga rumah, Metode *Random Forest*, *Principal Component Analysis*

Investasi menjadi hal yang menarik, khususnya investasi di bidang properti. Pihak developer juga harus berhati hati dalam menentukan harga properti. Perlu diketahui setiap tahunnya baik jangka pendek ataupun jangka panjang harga properti semakin naik dan bahkan hampir tidak pernah turun. Dalam menentukan harga sering juga berdasarkan dengan fitur yang dimiliki rumah seperti konsep, lokasi, kamar tidur, dll. Untuk memprediksi harga rumah berdasarkan fiturnya *random forest* mempunyai performa yang bagus untuk prediksi harga rumah. Namun metode *random forest* memiliki kelemahan jika penggunaan variabel terlalu banyak maka proses pelatihan menjadi lebih lama serta pemilihan fitur yang cenderung memilih fitur yang tidak informatif. Salah satu cara yang digunakan untuk mengurangi fitur tanpa harus menghapus fitur yang lain yaitu menggunakan *Principal Component Analysis*. Dalam penelitian ini motede yang digunakan adalah *Principal Component Analysis* dan *random forest*. Hasil pelatihan model dapat disimpulkan bahwa penggunaan hasil evaluasi model yang menggunakan PCA memiliki tingkat error yang lebih kecil dan nilainya lebih konsisten yaitu dengan rata-rata 0.0253. Sedangkan hasil evaluasi tanpa PCA dan hanya menggunakan *random forest* memiliki nilai eror yang lebih besar yaitu dengan rata rata 0.03275. Waktu pelatihan dengan menggunakan model PCA memiliki waktu yang lebih cepat dengan rata-rata 5007 milidetik, sedangkan yang hanya menggunakan *random forest* tanpa PCA memiliki waktu rata-rata 6099 milidetik.

ABSTRACT

Ahdan, Emha. 2023. **Optimization of the Random Forest Method Using Principal Component Analysis to Predict House Prices**. Theses. Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Promotor: (I) Dr. Totok Chamidy, M. Kom (II) Dr. Fresy Nugroho, M. T

keyword: House Price Prediction, Random Forest Method, Principal Component Analysis

Investment is an interesting thing, especially property investment. The developer must also be careful in determining the price of the property. It should be noted that every year, both short-term and long-term, property prices increase and rarely go down. In determining the price, it is often also based on the features of the house such as the concept, location, bedrooms, etc. To predict house prices based on their features, the random forest has a good performance for predicting house prices. However, the random forest method has the disadvantage that if you use too many variables, the training process will take longer and feature selection tends to select features that are not informative. One way to reduce features without removing other features is to use Principal Component Analysis. In this research, the method used is Principal Component Analysis (PCA) and Random Forest. From the results of model training, it can be concluded that the use of model evaluation results using PCA has a smaller error rate and more consistent values, with an average of 0.0253. While the results of the evaluation without PCA and using only Random Forest have a higher error value with an average of 0.03275. The training time using the PCA model has a faster time, with an average of 5007 miliseconds, while those using only random forest without PCA have an average time of 6099 miliseconds.

مستخلص البحث

احدا ، محمد. 2023. طريقة التحسين (Random Forest) باستخدام (Principal Component Analysis) للتنبؤ بأسعار المنازل. بحث جامعي. برنامج الماجستير قسم الهندسة والمعلوماتية، الجامعة الإسلامية الحكومية مولانا مالك إبراهيم بمالانج. المشرفة : (I) الدكتور توتوك شاميدي ،الماجستير (II) الدكتور فريزي نوجروهو،الماجستير

الكلمات المفتاحية تنبؤات أسعار المساكن,(Random Forest),(Principal Component Analysis)

الاستثمار شيء مثير للاهتمام ، وخاصة الاستثمار في العقارات. يجب على المطور أيضًا توخي الحذر في تحديد سعر العقار. وتجدر الإشارة إلى أنه في كل عام ، على المدى القصير والطويل ، تتزايد أسعار العقارات ولا تنخفض أبدًا. عند تحديد السعر ، غالبًا ما يعتمد أيضًا على ميزات المنزل مثل المفهوم والموقع وغرف النوم وما إلى ذلك. إن التنبؤ بأسعار المنازل بناءً على خصائصها (Random Forest) له أداء جيد في التنبؤ بأسعار المنازل. ومع ذلك ، فإن الطريقة (Random Forest) لها عيب أنه إذا كنت تستخدم الكثير من المتغيرات ، فستستغرق عملية التدريب وقتًا أطول ويميل اختيار الميزة إلى اختيار الميزات غير المفيدة. إحدى الطرق المستخدمة لتقليل الميزات دون الحاجة إلى إزالة الميزات الأخرى هي استخدام (Principal Component Analysis). الطريقة المستخدمة في هذا البحث هي (Principal Component Analysis) و (Random Forest). من نتائج تدريب النموذج ، يمكن استنتاج أن استخدام نتائج تقييم النموذج باستخدام (PCA) له معدل خطأ أقل وقيم أكثر اتساقًا ، بمتوسط 0.0253. في حين أن نتائج التقييم بدون (PCA) وباستخدام (Random Forest) فقط لها قيمة خطأ أعلى بمتوسط 0.03275. يتمتع وقت التدريب باستخدام نموذج PCA بوقت أسرع بمتوسط 5007 مللي ثانية ، في حين أن الوقت الذي يستخدم فقط (Random Forest) بدون (PCA) يبلغ متوسط وقت 6099 مللي ثانية.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Rumah merupakan merupakan istana bagi setiap muslim. Rumah menjadi tempat dimana banyak rahmat Alloh diturunkan kepada setiap muslim. Dalam Al-Qur'an surah Al-Araf ayat 74 yang berbunyi

وَاذْكُرُوا إِذْ جَعَلَكُمْ خُلَفَاءَ مِنْ بَعْدِ عَادٍ وَبَوَّأَكُمْ فِي الْأَرْضِ تَتَّخِذُونَ مِنْ سُهُولِهَا

قُصُورًا وَتَنْحِتُونَ الْجِبَالَ بُيُوتًا ۖ فَاذْكُرُوا آلَاءَ اللَّهِ وَلَا تَعْثَوْا فِي الْأَرْضِ مُفْسِدِينَ

Artinya: *"Dan ingatlah olehmu di waktu Tuhan menjadikam kamu pengganti-pengganti (yang berkuasa) sesudah kaum 'Aad dan memberikan tempat bagimu di bumi. Kamu dirikan istana-istana di tanah-tanahnya yang datar dan kamu pahat gunung-gunungnya untuk dijadikan rumah; maka ingatlah nikmat-nikmat Allah dan janganlah kamu merajalela di muka bumi membuat kerusakan." (QS. Al-Araf:74)*

Dalam Tafsir Jalalain dijelaskan bahwa Allah memberikan tempat bagi setiap muslim di bumi untuk didirikan istana (Rumah) dan Allah menyuruh untuk mengingat nikmat-nikmat Allah dan Allah menyuruh kita untuk tidak merusak apa yang disekitar kita.

Bisnis dibidang properti mempunyai keunggulan yang membuat setiap orang yang terjun didalamnya perlu hati hati. Terutama pada orang yang ingin berinvestasi di bidang properti. Memang investasi menjadi hal yang menarik, khususnya investasi di bidang properti. Sudah sejak tahun 2011 investasi properti ini meningkat secara signifikan baik secara on demand maupun secara penjualan (R.M.A. van der Schaar, 2015).

Pihak pengembang juga harus berhati hati dalam menentukan harga properti. Perlu diketahui setiap tahunnya baik jangka pendek ataupun jangka panjang harga properti semakin naik dan bahkan hampir tidak pernah turun (Feng & Jones, 2015). Salah memperhitungkan harga properti membuat bisnis dan investasi bisa menjadi rugi. Resiko ini sangatlah tidak bagus jika dialami oleh setiap pebisnis. Tentu saja setiap bisnis selalu mengejar yang namanya profit. Penting bagi seorang pebisnis, terutama dibidang properti untuk mengetahui dan mampu untuk memprediksi harga properti. Sehingga pebisnis dapat mendapatkan profit sebesar mungkin.

Diberbagai negara *House Price Index* (HPI) sering digunakan untuk menghitung kenaikan harga rumah (Garriga et al., 2021), namun kondisi fisik, konsep, lokasi juga mempengaruhi dalam menentukan harga rumah (Nur et al., 2017). Karakteristik fisik yang dimiliki rumah seperti jumlah kamar tidur, kamar mandi, ukuran bangunan dapat mempengaruhi harga rumah (Kang et al., 2021).

Ja'afar *et al.*, (2021) Melakukan penelitian *literature review* berkaitan dengan prediksi harga rumah. Ditemukan bahwa metode terbaik yang ditemukan adalah metode *random forest*. *Random forest* mempunyai performa yang bagus dari pada yang lain dalam kontek untuk prediksi rumah.

Metode *random forest* ini di buat oleh Leo Breimen (Breiman, 2001). Metode *random forest* ini bisa digunakan untuk menyelesaikan masalah klasifikasi (Gislason et al., 2004) ataupun masalah regresi (P. Jiang et al., 2007). Metode *random forest* ini ketika digunakan untuk melakukan prediksi harga rumah memiliki tingkat *error* yang rendah. Penelitian yang dilakukan oleh (Shahhosseini et al., 2020) tentang prediksi harga rumah dengan menggunakan dua jenis dataset

yaitu dataset rumah di Boston dan Ames memberikan hasil bahwa metode *random forest* memiliki tingkat error yang rendah yakni sebesar 0.0183.

Metode *random forest* ini memiliki kemampuan yang bagus untuk melakukan prediksi harga rumah. *Random forest* tergolong konsep pembelajaran *ensemble learning*, yaitu konsep yang merata-ratakan hasil dari beberapa pohon keputusan/*decission tree* yang diterapkan dari kumpulan data untuk meningkatkan akurasi. Dengan membuat jumlah pohon keputusan yang besar, metode ini dapat memberikan tingkat akurasi yang tinggi dan bisa menghindari masalah *overfitting*, namun hal itu mempunyai kelemahan yaitu meningkatkan waktu pelatihan yang lama (Adetunji et al., 2022). Selain itu pengacakan pada sampel *bagging* dan pemilihan fitur pada *random forest* cenderung memilih fitur yang tidak informatif untuk pemisahan node (Nguyen et al., 2015). Ini membuat *random forest* memiliki akurasi yang buruk saat bekerja dengan data dimensi tinggi.

Salah satu cara untuk mengurangi kekurangan tersebut yaitu dengan memilih fitur yang informatif, karena Pemilihan fitur merupakan langkah penting untuk mendapatkan performa yang baik untuk model (Nguyen et al., 2015). Selain itu pengurangan jumlah fitur dalam metode *random forest* dapat mempercepat kinerja dari model *random forest* tersebut. Salah satu cara yang digunakan untuk mengurangi fitur tanpa harus menghapus fitur yang lain yaitu dengan menggabungkan beberapa fitur yaitu menggunakan *Principal Component Analysis* (Gardner & Lo, 2021).

Principe Component Analysis atau PCA adalah sebuah teknik analisis data yang digunakan untuk melakukan pengurangan dimensi pada sebuah dataset

dengan tetap mempertahankan informasi pada dataset. PCA bekerja dengan mencari vektor utama (*principal components*) dari dataset yang memiliki variansi terbesar, yang kemudian digunakan untuk mewakili data. PCA umumnya digunakan ketika variabel dalam data memiliki kemiripan atau korelasi yang tinggi antar column

Penggunaan PCA dan *random forest* dapat meningkatkan kinerja menjadi semakin efektif, efisien dan dapat memberikan nilai akurasi yang tinggi dan error yang rendah (Lu et al., 2020; Song & Huang, 2021; Waskle et al., 2020). Lebih jelasnya sebagai berikut, Lu et al., (2020) melakukan penelitian mengenai Penilaian kerusakan struktur plastik yang diperkuat serat karbon (CFRP) dengan metode *random forest* dan PCA, hasilnya menunjukkan bahwa metode ini efektif. Song & Huang, (2021) melakukan penelitian untuk mengidentifikasi minuman keras palsu karena membahayakan kesehatan menggunakan metode *random forest* dan PCA hal ini mendapatkan akurasi yang tinggi yaitu sebesar 99,80%. Waskle et al., (2020) melakukan penelitian mengenai *intrusion detection system* (IDS) untuk membantu menemukan serangan pada sistem dan penyusup terdeteksi, metode yang digunakan yaitu metode *random forest* dan PCA, hasilnya menunjukkan bahwa nilai *performance time* sebesar 3,24 menit, Tingkat Akurasi sebesar 96,78 %, dan tingkat error sebesar 0,21%.

Dari pemaparan sebelumnya maka perlu dilakukan sebuah penelitian yang digunakan untuk mengembangkan sebuah sistem yang nantinya dapat membantu para pebisnis properti untuk memprediksi harga properti yang digunakan untuk memaksimalkan profit. Perlu sebuah metode untuk melakukan prediksi harga

berdasarkan fitur fitur yang dimiliki oleh properti tersebut. Dalam penelitian ini motede yang digunakan adalah *principal component analysis* dan *random forest*.

1.2 Pernyataan Masalah

Dari latar belakang yang telah dibahas, rumusan masalah yang akan diangkat yaitu

1. Seberapa besar tingkat *error* apabila *metode Principal Component Analysis* dan *Random Forest* digunakan untuk melakukan prediksi harga rumah.
2. Metode apa yang lebih optimal untuk memprediksi harga rumah antara metode *random forest* yang sudah di optimalkan menggunakan metode *Principal Component Analysis* dan yang belum dioptimalkan.

1.3 Tujuan Penelitian

Dari rumusan masalah yang telah dibahas maka tujuan dari penelitian ini yaitu untuk

1. Mengukur seberapa besar tingkat error apabila metode *Principal Component Analysis* dan metode *Random Forest* digunakan untuk melakukan prediksi harga rumah.
2. Membandingkan antara metode *Random Forest* yang sudah di optimalkan menggunakan metode *Principal Component Analysis* dan yang belum dioptimalkan.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini diharapkan mampu membantu para pebisnis dan investor properti untuk bisa meningkatkat keuntungan dan mengurangi resiko terjadi kerugian dalam bisnis properti.

1.5 Batasan Masalah

Batasan masalah dalam penelitian ini untuk membatasi penelitian agar tidak melebar kemana mana yaitu

1. Penelitian ini akan menyelesaikan masalah regresi.
2. Tingkat error digunakan untuk mengukur kinerja model.
3. Penentuan harga rumah berdasarkan dari fitur yang ada dalam rumah atau *hedonic pricing.*
4. Data yang digunakan adalah harga rumah kota malang dari hasil *scraping* data di website rumah.com yang dilakukan pada tanggal 17 februari 2023.
5. Penelitian ini sampai tahap evaluasi untuk melihat tingkat *error* pada model, dan tidak sampai tahap *deployment* dan pemberian *feedback*.

BAB II
LITERRATURE REVIEW

2.1 House Price Prediction

Lim *et al.*, (2016) Melakukan studi prediksi harga rumah dengan membandingkan dua metode yaitu metode ANN dan ARIMA. Penting bagi calon pembeli untuk mengambil keputusan, sehingga diperlukan penelitian untuk mendapatkan metode terbaik dalam membuat prediksi yang akan menguntungkan pemangku kepentingan. Dataset yang digunakan adalah data perumahan di Singapura. Dari hasil penelitian diperoleh hasil sebagai berikut untuk metode JST memiliki nilai MAE 1.9, RMSE 3.1 MAPE 1.7%, dan untuk metode ARIMA memiliki nilai MAE 3.2 RMSE 5.0 MAPE 2.5%. Metode unggulan dengan nilai MSE terendah adalah JST.

De Nadai & Lepri (2018) Melakukan penelitian tentang prediksi harga rumah. Penelitian juga menjelaskan beberapa data yang dapat mempengaruhi seperti kontribusi ekonomi dari karakteristik lingkungan seperti walkability dan persepsi keamanan. Metode yang digunakan dalam penelitian XGBoost, *Gradient Boosted Trees*. Eksperimen yang melibatkan 70.000 rumah di 8 kota Italia menyoroti bahwa vitalitas dan kemudahan berjalan kaki di lingkungan tersebut tampaknya mendorong lebih dari 20% nilai perumahan. Selain itu, penggunaan informasi ini meningkatkan siaran saat ini sebesar 60%. Oleh karena itu, penggunaan karakteristik lingkungan suatu properti dapat menjadi sumber daya yang tak ternilai untuk menilai nilai ekonomi dan sosial rumah setelah perubahan lingkungan dan, secara potensial, mengantisipasi gentrifikasi. Parameter yang

digunakan untuk studi model adalah MAE MdAPE. Parameter ini digunakan untuk memprediksi harga *real estat* perumahan untuk tiga model berbeda. Yang pertama adalah properti hanya menggunakan fitur tekstual dari rumah (misalnya jumlah kamar, nomor lantai). Model kedua menggunakan fitur tekstual dan *ego-place*. Ketiga versi Open tersebut hanya menggunakan data kontekstual dengan lisensi Open. Berikut hasil yang didapatkan model pertama didapat MAE148, 109 MdAPE 23,78%, model kedua MAE 104, 586 MdAPE 15,44%, MAE 138, 929 MdAPE 18,02%.

Phan (2018) Melakukan penelitian prediksi harga rumah dengan menggunakan kombinasi algoritma *machine learning*. Data yang digunakan adalah data asli tentang pasar perumahan yang terdiri dari 34.857 data dan 21 variabel yang diunduh dari Kaggle, algoritma yang digunakan dalam memprediksi adalah kombinasi dari algoritma *Stepwise* dan *Support Vector Machine*, serta kombinasi dari model *Stepwise* dan SVM adalah pendekatan kompetitif dengan nilai MSE data pelatihan diperoleh nilai 0,0558 dan data pengujian diperoleh MSE 0,0615, dengan rasio evaluasi 0,62. Dari hasil penelitian terungkap bahwa terdapat perbedaan harga yang tinggi antara harga rumah termahal di pinggiran kota dengan yang paling terjangkau di kota Melbourne.

Durganjali & Pujitha (2019) Melakukan penelitian prediksi harga rumah dengan menggunakan beberapa algoritma klasifikasi yang berbeda seperti Regresi Logistik, *Decision tree*, Naive Bayes, dan *random forest* dan dalam penelitiannya juga menggunakan algoritma *AdaBoost* untuk meningkatkan proses pembelajaran mesin untuk hasil yang lebih baik. Data yang digunakan diperoleh dari situs penyedia data yaitu Kaggle yang memiliki 217006 baris dan 13 kolom. Setelah

dilakukan proses pembelajaran pada model, diperoleh hasil regresi logistik dengan akurasi 81,5%, *Decission Tree* dengan akurasi 91%, *random forest* dengan akurasi 86,5% naive bayes dengan akurasi 88%, dan terdapat peningkatan sebesar 96 % akurasi. Sehingga dari hasil tersebut diperoleh bahwa *Adabosst* dan *Decission Tree* memiliki nilai akurasi yang tinggi.

Jiang & Shen (2019) Melakukan penelitian prediksi harga rumah dengan menggunakan dataset harga rumah bekas di Shanghai. Ada 61609 baris dan 29 kolom dalam dataset. Penelitian yang dilakukan oleh Jiang & Shen, 2019 menggunakan model *Back Propagation Neural Network* yang diimplementasikan dengan *framework Keras Deep Learning Framework*. setelah dilakukan pengujian diperoleh hasil dengan akurasi yang tinggi, dengan nilai akurasi sebesar 95,59%.

Madhuri *et al.* (2019) Melakukan penelitian prediksi harga rumah dengan menggunakan teknik regresi. Di dalamnya terdapat beberapa metode yang digunakan untuk membandingkan metode mana yang memiliki akurasi lebih baik dari beberapa metode yang telah dijelaskan. Peneliti bertujuan untuk memberikan bantuan mengenai prediksi harga rumah bagi setiap orang yang ingin membeli rumah. Dalam penelitian metode yang digunakan adalah metode *Multiple linear*, Ridge, LASSO, Elastic Net, *Gradient boosting*, dan *AdaBoost Regression*. algoritma *gradient boosting* memiliki nilai akurasi yang tinggi jika dibandingkan dengan semua algoritma lain mengenai prediksi harga rumah. kumpulan data publik. Dari penelitian ini didapatkan bahwa algoritma *gradient boosting regression* memiliki skor yang baik dengan skor error 0.917.

Shahhosseini *et al.* (2020) Dilakukan penelitian prediksi harga rumah dengan menggunakan dua jenis dataset yaitu dataset rumah di Boston dan Ames.

penelitiannya menggunakan tujuh algoritma yang digunakan dalam pembuatan model machine learning yaitu 1. *regresi LASSO* 2. *Random Forests 3. Deep Neural Network 4. Extreme Gradient Boosting (XGBoost) 5. Support Vector Machines dengan kernel polynomial 6. Support Vector Machines* dengan kernel RBF 7. Mendukung Mesin Vektor dengan kernel sigmoid. Hasil dari algoritma ini adalah dataset perumahan di Boston dengan MSE terendah adalah XGBoost, dengan skor MSE sebesar 21.367. Sedangkan pada dataset perumahan di Ames Ames yang memiliki MSE terendah adalah algoritma LASSO dengan skor 0,0132.

Sharma *et al.* (2021) dalam penelitiannya berfokus pada analisis data, menjelaskan tahapan analisis data. Teknik yang digunakan adalah teknik *Exploratory Data Analytics* (EDA). obyek penelitian tentang prediksi nilai rumah di negara bagian California. Ini termasuk informasi mentah dan data tentang rumah seperti — garis bujur, garis lintang, jumlah kamar, dan sebagainya. Model yang digunakan dalam penelitian ini adalah *Decission Tree* dan *Random Forest* dan memiliki akurasi 80%.

Adetunji *et al.* (2022) Melakukan penelitian prediksi harga rumah. Pada penelitian ini fokus pada penyelesaian masalah dengan 1 metode yaitu metode *random forest*. Data yang digunakan adalah dataset rumah di Boston yang memiliki 506 data dan 14 kolom/fitur. Ditampilkan juga heatmap yang menjelaskan hubungan antar fitur, fitur mana yang mempengaruhi fitur lainnya. Diketahui bahwa fitur tax dan RAD merupakan fitur yang sangat berpengaruh. Kemudian setelah dilakukan pengujian model, diperoleh hasil matriks berikut untuk *r-squere* 0.9, MAE 1.9, MSE 6.7, RMSE 2.5.

Wiradinata *et al.* (2022) Melakukan riset dengan terobosan baru dengan memberikan gambaran harga rumah pasca COVID-19 di wilayah Surabaya Indonesia. Penelitian ini membandingkan 3 metode untuk menghasilkan hasil terbaik. Metode-metode tersebut adalah Regresi Linier, Regresi *Decission Tree, Regresi Random Forest*. Dataset yang digunakan diperoleh dari lima agen properti di Surabaya dan juga web *scraping* dari portal penjualan rumah online. Hasil prediksi harga rumah terbaik yang diperoleh dari penelitian ini adalah dengan menggunakan metode *random forest* dengan nilai RMSE 885,3, MAPE 0,23, R Squere 0,8.

2.2 Theoritical Framework

Pada bagian ini akan dibahas beberapa teori yang mendukung penelitian tentang prediksi harga rumah. Ada beberapa metode terbaik yang akan dipilih yang akan digunakan dalam penelitian ini. Berikut ini adalah kerangka teoritis untuk memprediksi harga rumah.

Tabel 2. 1 Theoritical Framwork for House Price Prediction

<table>
<tr><th>Input</th><th>Proses</th><th>Output</th></tr>
<tr><td>• Harga Rumah
• Fitur Rumah</td><td>
<table>
<tr><td>Madhuri et al., 2019</td><td>gradient boosting</td><td>Weight</td><td>physical conditions, concept and location etc</td><td>0.917</td></tr>
<tr><td>Phan, 2018</td><td>SVM</td><td>Weight</td><td>price, Property Count, year,Distance, Longitude, Latitude,Rooms,Bathrooms,Car,Land size, Type</td><td>0.0615</td></tr>
<tr><td>Lim et al., 2016</td><td>ANN</td><td>Weight</td><td>CAP,Bedroom, Bathroom,Floor Area,Tenure,Age,MRT,School,Shopping,Mall,Childcare,Centre</td><td>1.9</td></tr>
<tr><td>Lim et al., 2016</td><td>ARIMA</td><td>Weight</td><td>CAP,Bedroom, Bathroom,Floor Area,Tenure,Age,MRT,School,Shopping,Mall,Childcare,Centre</td><td>3.2</td></tr>
<tr><td>Jiang & Shen, 2019</td><td>BP Neural Network</td><td>Weight</td><td>kitchen, room, saloon. coordinat, etc</td><td>4,410</td></tr>
<tr><td>Durganjali & Pujitha, 2019</td><td>decision tree</td><td>Tree</td><td>house price, storage range, floor size details, etc.</td><td>9,000</td></tr>
<tr><td>Sharma et al., 2021</td><td>decision tree</td><td>Tree</td><td>longitude,latitude,house median age, total rooms, total bedrooms, population, house holds, median income,median house value</td><td>20,000</td></tr>
<tr><td>Sharma et al., 2021</td><td>random forest</td><td>Tree</td><td>longitude,latitude,house median age, total rooms, total bedrooms, population, house holds, median income,median house value</td><td>20,000</td></tr>
<tr><td>De Nadai & Lepri, 2018</td><td>XGBoost</td><td>Tree</td><td>square meters, built year, energy certification, monthly expenses (condominium), floor number, heating type, type of fixtures, garden, furnished, terrace, sun exposition, kitchen type, spa, cellar, garage, fireplace, place type, property class and type, property taxes, condition, and number of rooms, bathrooms, bedrooms</td><td>15.44</td></tr>
<tr><td>Shahhosseini et al., 2020</td><td>XGBoost</td><td>Tree</td><td>YearBuilt, Neighborhood,Street,BldgType,MSSubClass, Foundation, LotArea, RoofStyle, Bedroom, FullBath, TotalBsmtSF, 1stFlrSF, TotRmsAbvGrd, GrLivArea, GarageCars, GarageArea, OverallQual, ExterQual, KitchenQual, BsmtQual, SalePrice</td><td>21.367</td></tr>
<tr><td>Adetunji et al., 2022</td><td>random forest</td><td>Tree</td><td>Crime, ZN, River, Industri, Nox, Room,Age, Distance to town, Tax, Radial Highways, Ratio student-teacher, proportion of blacks by town, lower status of the population, Price</td><td>1.9</td></tr>
<tr><td>Wiradinata et al., 2022</td><td>random forest</td><td>Tree</td><td>Cluster name, Surface area, Building area, Bedrooms,Bathrooms, Storey, Ownership status, Facing, House position, Road width, Urgent, Building age, Ready to use, Furnished, Category, Pricing category, Community price,Price</td><td>0.23</td></tr>
<tr><td>Shahhosseini et al., 2020</td><td>random forest</td><td>Tree</td><td>YearBuilt, Neighborhood,Street,BldgType,MSSubClass, Foundation, LotArea, RoofStyle, Bedroom, FullBath, TotalBsmtSF, 1stFlrSF, TotRmsAbvGrd, GrLivArea, GarageCars, GarageArea, OverallQual, ExterQual, KitchenQual, BsmtQual, SalePrice</td><td>0.0183</td></tr>
</table>
</td><td>Prediksi Harga Rumah</td></tr>
</table>

Variabel input memengaruhi kinerja dalam algoritma pembelajaran mesin. Dalam hal ini yang diinputkan dalam beberapa makalah adalah harga rumah dan ciri-ciri rumah tersebut. Banyak algoritma yang dapat diterapkan untuk mendapatkan metode yang baik, penulis merangkum beberapa metode yang didapat

dari beberapa paper mengenai prediksi harga rumah. kemudian penulis mereview beberapa algoritma yang telah didapatkan, mencari nilai akurasi yang tinggi dan nilai error yang rendah. Penulis telah merangkumnya dalam tabel berikut.

Tabel 2. 2 Best Performing Method

Penulis	**Metode**		**Matrik**	
Shahhosseini et al., 2020	*random forest*	Tree	Error	0,018
Phan, 2018	SVM	Weight	Error	0,062
Wiradinata et al., 2022	*random forest*	Tree	Error	0,230
Madhuri et al., 2019	gradient boosting	Weight	Error	0,917
Lim et al., 2016	ANN	Weight	Error	1,900
Adetunji et al., 2022	*random forest*	Tree	Error	1,900
Lim et al., 2016	ARIMA	Weight	Error	3,200
Durganjali & Pujitha, 2019	Adabosst	Tree	Error	4,000
Jiang & Shen, 2019	BP Neural Network	Weight	Error	4,410
Durganjali & Pujitha, 2019	decision tree	Tree	Error	9,000
De Nadai & Lepri, 2018	XGBoost	Tree	Error	15.440
Sharma et al., 2021	decision tree	Tree	Error	20,000
Sharma et al., 2021	*random forest*	Tree	Error	20,000
Shahhosseini et al., 2020	XGBoost	Tree	Error	21,367

Pada tabel terlihat bahwa terdapat *random forest* memiliki tingkat kesalahan yang paling kecil yaitu 0,0183. Dari studi literatur, metode *random forest* merupakan metode yang sering digunakan dan memiliki kinerja yang lebih baik.

2.3 Optimasi

Optimalisasi adalah proses menemukan solusi terbaik atau kondisi optimal dalam konteks tertentu. Algoritma dan teknik matematika digunakan untuk meningkatkan efisiensi, efektivitas, atau profitabilitas sistem. Optimasi memiliki banyak bidang dan aplikasi termasuk optimasi matematis, optimasi numerik, optimasi kombinatorial, optimasi stokastik, dll.

2.4 *Principal Component Analysis*

Principe Component Analysis atau PCA adalah sebuah teknik analisis data yang digunakan untuk melakukan pengurangan dimensi pada sebuah dataset dengan tetap mempertahankan informasi pada dataset. PCA bekerja dengan mencari vektor utama (*principal components*) dari dataset yang memiliki variansi terbesar, yang kemudian digunakan untuk mewakili data.

Fitur PCA yang paling berguna adalah menghasilkan transformasi linear dari data yang dihitung dengan perkalian matriks sederhana. Ini sangat berguna dalam skenario pengurangan dimensi ketika menghitung fungsi kompleks dari seluruh kumpulan data akan sangat mahal dalam hal waktu komputasi (Smallman *et al*., 2018).

Untuk data yang bertipe tekstual biasanya diubah menjadi data numerik sehingga alat statistik klasik dapat diterapkan padanya. Salah satu cara paling umum untuk melakukannya adalah dengan membuat "Matriks Istilah Dokumen" di mana setiap baris mewakili pengamatan dan setiap kata/istilah unik dalam koleksi mewakili kolom (Smallman et al., 2018).

Penelitian lain yang dilakukan oleh El Boujnouni *et al*., (2021), melakukan juga reduksi dimensi terhadap dataset yang bertipe textual. Tulisan ini bertujuan untuk mengetahui asal usul virus ini dengan cara membandingkan urutan asam nukleatnya dengan semua anggota famili coronaviridae. Sebelum melaukan reduksi dimensi menggunakan PCA dilakukan konversi data text menggunakan *Ngrams analysis of genomes*.

2.5 Proses *Principal Component Anlysis*

Principe Component Analysis atau PCA adalah sebuah teknik analisis data yang digunakan untuk melakukan pengurangan dimensi pada sebuah dataset dengan tetap mempertahankan informasi pada dataset. PCA bekerja dengan mencari vektor utama (*principal components*) dari dataset yang memiliki variansi terbesar, yang kemudian digunakan untuk mewakili data. PCA umumnya digunakan ketika variabel dalam data memiliki kemiripan atau korelasi yang tinggi antara column. Kemiripan ini sering disebutnya data yang berulang atau redundant.

Berikut adalah langkah-langkah umum dalam melakukan PCA:

1. Normalisasi data: Data harus dalam bentuk skala yang sama agar perbandingan antar fitur tidak terpengaruh oleh skalanya.
2. Menghitung matrik kovariansi: Matrik kovariansi digunakan untuk menghitung variansi dan korelasi antar fitur dalam dataset. Setelah diketahui varians dan kovariannya langkah selanjutnya yaitu menghitung matrik kovariannya
3. Menghitung *eigenvectors* dan *eigenvalue*s: *Eigenvectors* adalah vektor utama yang memiliki variansi terbesar dan *eigenvalue*s adalah variansi yang terkait dengan masing-masing *eigenvector*. *Eigenvectors* dan *eigenvalue*s dari matrik kovariansi C dapat ditemukan dengan memecahkan persamaan *eigenvalue* berikut $Cv = \lambda v$ di mana λ adalah *eigenvalue* dan v adalah *eigenvector*
4. Pemilihan *eigenvectors*: *Eigenvectors* dengan variansi terbesar dipilih sebagai principal components.

5. Proyeksi data: Data dapat dikompresi dengan memproyeksikan data ke dalam subruang spasial yang didefinisikan oleh principal components.

2.6 *Principal Component Analysis* and *Random Forest*

Čeh et al., (2018) melakukan penelitian mengenai prediksi harga apartemen menggunakan metode *random forest* yang dikombinasikan dengan metode *principel component analysis*. Nantinya hasil kinerja dari metode *random forest* ini dibandingkan dengan model hedonis yang umum digunakan berdasarkan regresi berganda untuk prediksi harga apartemen. Kumpulan data yang mencakup 7407 catatan transaksi apartemen yang mengacu pada penjualan *real estat* dari 2008-2013 di kota Ljubljana, ibu kota Slovenia. Dari hasil penelitian diperoleh bahwa *random forest* dan PCA menunjukkan hasil jauh lebih baik untuk memprediksi dengan nilai MAPE atau error sebesar 7,27%.

Gardner & Lo, (2021) Dipenelitian lain meneliti mengenai arsitektur baru untuk algoritma *random forest*. Arsitektur ini berupaya meningkatkan kemampuan *Random Forest* untuk mengenali interdependensi fitur. Peningkatan kinerja dicapai dengan membuat model PCA di dalam setiap pohon. Model PCA ini membuat fitur tambahan baru yang berisi informasi dari beberapa fitur masukan. Fitur-fitur baru ini tidak digunakan untuk pengurangan fitur melainkan ditambahkan ke vektor fitur yang ada. Dari hasil penelitian menghasilkan model yang mendukung presisi tinggi dengan skor F1 yang hampir setara dengan algoritma *random forest* tradisional.

Waskle et al., (2020) melakukan penelitian mengenai intrusion detection system (IDS) untuk membantu menemukan serangan pada sistem dan penyusup terdeteksi, metode yang digunakan yaitu metode *random forest* dan PCA, hasilnya

menunjukkan bahwa pendekatan yang diusulkan bekerja lebih efisien dalam hal akurasi dibandingkan dengan teknik lain seperti SVM, Naïve Bayes, dan Decision Tree dengan nilai performance time sebesar 3,24 menit, Tingkat Akurasi sebesar 96,78 %, dan Tingkat Error sebesar 0,21%.

Lu et al., (2020) melakukan penelitian mengenai Penilaian kerusakan struktur plastik yang diperkuat serat karbon (CFRP) yang akurat dan tepat waktu sangat penting untuk memastikan keamanan struktur komposit penerbangan dan kedirgantaraan dengan metode *random forest* dan PCA, hasilnya menunjukkan bahwa metode ini efektif.

Song & Huang, (2021) melakukan penelitian untuk mengidentifikasi minuman keras palsu karena membahayakan kesehatan menggunakan metode *random forest* dan PCA. sangat penting bagi kesehatan manusia dan perlindungan merek. Makalah ini menyajikan cairan pengenal fluoresensi yang diinduksi laser (yang memiliki sensitivitas dan kecepatan sangat tinggi). Secara khusus, pertama-tama ia menyiapkan sistem akuisisi spektrum LIF untuk minuman keras di laboratorium, menggunakan Analisis Komponen Utama (PCA) untuk mengurangi dimensi data spektral LIF, dan kemudian mengumpulkan data pengurangan dimensi untuk identifikasi menggunakan *random forest*. Hal ini mendapatkan akurasi yang tinggi yaitu sebesar 99,80%.

Dari beberapa literature review mengenai penelitian yang menggunakan metode *random forest* dan *principal component analysis* berikut adalah rangkuman dari literature review.

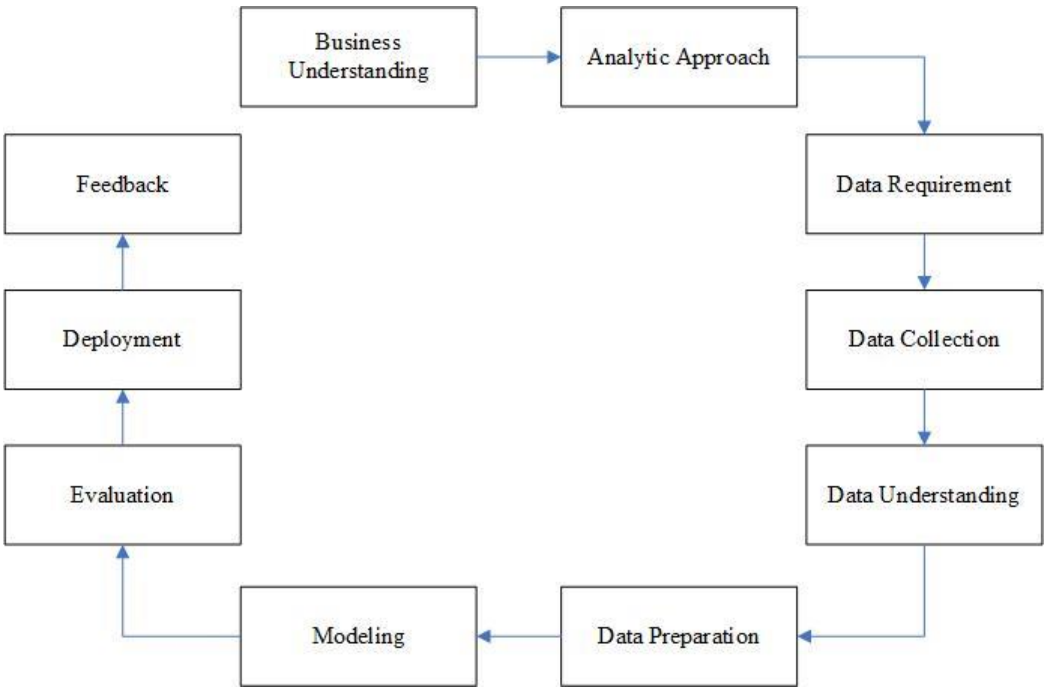
Tabel 2. 3 Rangkuman literature review *random forest* dan *principal component analysis*

Penulis	**Dataset**	**Matrik Penilaian**	**Nilai**
Čeh et al., (2018)	Transaksi penjualan apartemen di Ljubljana, ibu kota Slovenia	MAPE	7,27%.
Gardner & Lo, (2021)	PaySim Dataset (transaksi uang seluler sesuai dengan kumpulan data asli)	F1-Score	0.892
Waskle et al., (2020)	KDD Dataset	Error	0,21%.
Lu et al., (2020)	Carbon fiber reinforced plastics (CFRP) dataset	Akurasi	Efektif
Song & Huang, (2021)	data spektral LIF	Akurasi	99,80%

BAB III
METODOLOGI PENELITIAN

3.1 Desain Penelitian

Desain sistem ini menggambarkan bagaimana sistem yang akan dibangun bekerja. Diawali dengan *business understanding*, *analytic approach*, *data requirements*, *data collection*, ditahap ini tahap pengambilan data penjualan rumah kota malang yang diambilkan dari hasil *scraping* website rumah.com dan pemahan data atau *data understanding*. Setelah pengumpulan data tahap selanjutnya adalah tahap *data preparation* dimana *principal component analysis* diterapkan. Kemudian untuk perencanaan sistem menggunakan *modeling* dengan metode *random forest*. Selanjutnya adalah melihat hasil evaluasi dan menganalisisnya dan mengambil kesimpulan. Dalam penelitian ini sudah dibatasi dalam batasan masalah sampai tahap *evaluation*. Alur peoses penelitian dapat dilihat di gambar 3.1



Gambar 3. 1 Desain Penelitian

3.1.1 *Business Understanding*

Dalam Kasus ini perlu didefinisikan masalah dari proses *data mining* untuk mengetahui dalam proses ini apa yang ingin diselesaikan. Setelah diketahui masalah maka langkah selanjutnya yaitu mengetahui tujuan yang ingin dicapai. Serta mengidentifikasi sumber daya yang dibutuhkan.

Selain memahami mengenai masalah dan tujuan, dalam tahap ini juga diidentifikasi juga metode yang akan digunakan. Setelah mengidentifikasi metode yang digunakan selanjutnya mengidentifikasi metrik yang digunakan untuk mengevaluasi seberapa baik model dalam melakukan prediksi harga rumah.

3.1.2 *Analytic Approach*

Setelah masalah bisnis dinyatakan dengan jelas, langkah selanjutnya yaitu menentukan pendekatan analitis untuk memecahkan masalah tersebut. Langkah ini melibatkan pemaparan masalah dalam konteks teknik statistik dan *machine learning*, dan penting karena membantu mengidentifikasi model apa yang diperlukan untuk menjawab pertanyaan dengan paling efektif.

3.1.3 *Data Requirement*

Tahap adalah tahap di mana mengidentifikasi konten, format, dan sumber data yang diperlukan untuk pengumpulan data awal yang akan digunakan dalam penelitan.

3.1.4 *Data Collection*

Langkah ini mengidentifikasi sumber daya data yang tersedia yang relevan dengan masalah. Untuk mengambil data, kita dapat melakukan pengumpulan web

dari situs web terkait atau menggunakan repositori kumpulan data yang dibuat sebelumnya. Biasanya, kumpulan data siap adalah file CSV atau Excel.

3.1.5 *Data Understanding*

Dalam tahap ini dilakukan untuk memahami data yang akan digunakan untuk melatih model. Tahap ini memungkinkan untuk memahami informasi dalam data dan menentukan kualitasnya.

Pengumpulan data pada penelitian ini menggunakan data publik, diambil dari hasil *scraping* dari website penjualan rumah kota malang yaitu dari website rumah.com. Didalam dataset tersebut terdapat 6130 baris dan 11 kolom. Berikut adalah tabel penjelasan dari variabel, type, tipe data dan deskripsi.

Tabel 3. 1 Tabel Variabel Dataset dan Deskripsi

No	Variabel	Type	Tipe Data	Deskripsi
1	jumlah kamar tidur	Fitur Rumah	float	Jumlah kamar tidur di rumah untuk dijual
2	jumlah kamar mandi	Fitur Rumah	float	Luas rumah dalam meter persegi
3	luas tanah	Fitur Rumah	float	Luas rumah dalam meter persegi
4	harga per meter	Fitur Rumah	float	Nilai harga sebuah rumah dihitung per meter persegi
5	alamat	Fitur Rumah	Object	Lokasi rumah yang dijual
6	luas bangunan	Fitur Rumah	float	Luas bangunan dalam meter persegi
7	Sertifikat	Fitur Rumah	Object	Sertifikat penjualan rumah
8	Interior	Fitur Rumah	Object	Jenis interior yang ada di dalam rumah
9	parkir	Fitur Rumah	float	Jumlah tempat parkir di rumah
10	listrik	Fitur Rumah	float	Tegangan listrik di rumah dijual dalam satuan watt
11	harga	Target	float	Nilai harga rumah yang dijual

Berikut adalah contoh data yang akan digunakan dalam penelitian ini

kamar tidur	kamar mandi	luas tanah(m2)	harga per m	Alamat	luas Bangunan(m2)	sertifikat	interior	parkir	listrik	harga
6	6	291	10.309278	Blimbing	450	SHM - Sertifikat Hak Milik	Sebagian	1.0	2200.0	3.000000e+09
2	2	85	8.823529	Sukun	60	SHM - Sertifikat Hak Milik	Tak Berperabot	1.0	1300.0	7.500000e+08
10	10	118	27.542373	Lowokwaru	300	SHM - Sertifikat Hak Milik	Lengkap	1.0	2200.0	3.250000e+09
10	10	160	21.875000	Lowokwaru	350	SHM - Sertifikat Hak Milik	Sebagian	1.0	2200.0	3.500000e+09
10	10	149	19.463087	Lowokwaru	300	SHM - Sertifikat Hak Milik	Lengkap	1.0	2200.0	2.900000e+09

Gambar 3. 2 Dataset

Ada beberapa hal yang dapat dilakukan dalam tahap ini *data understanding*, yaitu

a. *Exploratory Data Analysis - Data Loading*

Data loading adalah proses pengambilan atau pemuatan data dari sumber eksternal ke sistem atau aplikasi yang sedang digunakan. Data dapat diambil dari berbagai sumber seperti database, file, atau server eksternal.

b. *Exploratory Data Analysis* - Deskripsi Variabel

Deskripsi variabel adalah proses mengidentifikasi dan menggambarkan sifat-sifat dari data yang diterima oleh variabel tersebut. Deskripsi variabel meliputi informasi seperti tipe data, rentang nilai, distribusi, dan variasi.

c. *Exploratory Data Analysis* - Menangani *Outlier* dan *Missing Value*

Outlier adalah nilai yang sangat berbeda dari kebanyakan data dalam suatu variabel. *Missing value* adalah data yang tidak tersedia atau tidak tercatat dalam suatu variabel. Menangani *Outlier dan Missing Value* sangat penting karena kedua hal ini dapat mempengaruhi hasil analisis data.

Ada beberapa cara yang digunakan untuk mendeteksi *outlier*, salah satunya menggunakan teknik IQR method. IQR adalah singkatan dari *Inter Quartile Range*.

IQR method (*Interquartile Range method*) adalah salah satu teknik untuk menentukan apakah suatu nilai adalah *outlier* atau tidak. IQR didasarkan pada konsep bahwa nilai yang berada diluar rentang interquartil (Q3-Q1) dapat dikategorikan sebagai *outlier*. Rumus IQR adalah sebagai berikut:

$$IQR = Q3 - Q1 \tag{3.1}$$

Langkah-langkah untuk menggunakan IQR method dalam mendeteksi *outlier* sebagai berikut

1. Hitung Q1 dan Q3: Q1 adalah nilai pada kuartil 1 (25%) dan Q3 adalah nilai pada kuartil 3 (75%) dari data.
2. Hitung IQR: Dengan mengurangi Q1 dari Q3, sehingga akan didapat nilai IQR.
3. Tentukan Batas Bawah dan Batas Atas: Batas bawah dapat ditemukan dengan mengurangi IQR dengan Q1, sedangkan batas atas dapat ditemukan dengan menambahkan IQR pada Q3.
4. Bandingkan Nilai Data dengan Batas Bawah dan Batas Atas: Jika suatu nilai data berada diluar batas bawah dan batas atas, maka nilai tersebut dapat dikategorikan sebagai *outlier*.

Setelah diketahui *outlier* dalam data, langkah yang akan dilakukan yaitu menghapus row yang mengandung *outlier* tersebut. Sehingga data terbebas dari *outlier* sebelum masuk ke tahap pelatihan model.

d. Exploratory Data Analysis - Univariate Analysis

Univariate Analysis adalah teknik analisis data yang memfokuskan pada satu variabel saja. Tujuannya adalah untuk memahami karakteristik dari suatu variabel, seperti distribusi, *central tendency*, dan variabilitas. Univariate Analysis meliputi tipe-tipe analisis seperti deskripsi statistik, histogram, dan box plot. *Univariate Analysis* juga dapat digunakan sebagai langkah awal dalam analisis data untuk menentukan apakah suatu variabel memiliki kaitan yang signifikan dengan variabel lain.

e. Exploratory Data Analysis - Multivariate Analysis

Multivariate Analysis adalah teknik analisis data yang memfokuskan pada lebih dari satu variabel sekaligus. Tujuannya adalah untuk memahami hubungan antar variabel dan bagaimana satu variabel mempengaruhi variabel lain. *Multivariate Analysis* sangat berguna dalam situasi di mana seseorang ingin memahami bagaimana variabel-variabel berinteraksi dan mempengaruhi suatu konsekuensi.

3.1.6 *Data Preparation*

Tahap *data preparation* dilakukan untuk melakukan proses pra premrosesan data sebelum data masuk ke tahap modeling. Tahap ini penting karena dalam tahap ini data akan di tranformasikan menjadi data yang cocok untuk proses modeling. Adapun hal-hal yang akan dilakukan dalam tahap *data preparation* sebagai berikut

1. Data Preparation - Data Encoding

Tahap ini adalah tahap dimana data yang berbentuk kategorik akan dirubah menjadi data numerik. Dibeberapa kasus data yang bersifat kategorik tidak bisa diproses oleh algoritma machine learning. Dalam

ada beberapa cara yang bisa dilakukan untuk mengubah data kategorik menjadi data numerik yaitu menggunakan teknik one-hot encoding.

One-Hot encoding adalah salah satu cara yang dilakukan untuk mengubah data kategorik menjadi data numerik. Dalam tahap ini setiap data kategorik akan dikodekan menjadi fitur dummy yang dirubah menjadi nol atau satu.

2. *Data Preparation* - Reduksi Dimensi dengan PCA

   Tahap ini menjadi tahap bahasan khusus dalam penelitian. Dimana dalam tahap ini akan dibahas khusus pada sub-bab tersendiri agar pembahasan menjadi lebih rinci.

3. *Data Preparation* - *Split Data Training* dan *Testing*

   Tahap split data ini dilakukan untuk membagi jumlah dataset yang dimiliki menjadi data *training* dan data *testing*. Data *training* ini yang nantinya digunakan untuk melatih model dan data *testing* digunakan untuk melakukan tes terhadap model yang sudah dilatih.

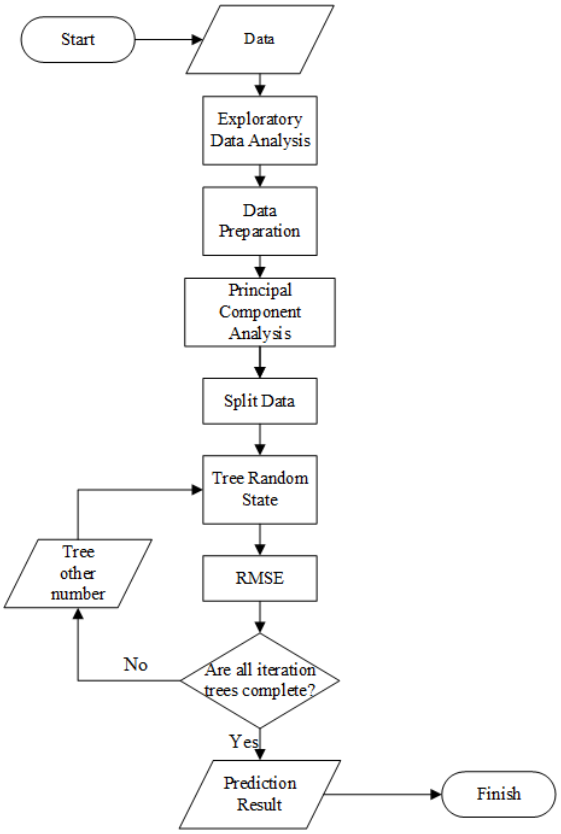
   Tahap split data ini dilakukan sebelum dilakukannya proses standarisasi pada data. Perlu mempertahankan sebagian data yang ada dalam kontek ini adalah data *testing* untuk menguji seberapa generalisasi model dalam membaca data baru sebelum di normalisasi atau di standarisasi (Fuentes, 2018).

4. *Data Preparation* - Normalisasi

   Algoritma *machine learning* akan mempunyai performa yang baik apabila data yang digunakan memiliki skala yang hampir sama, makanya dalam hal ini akan dilakukan standarisasi pada data. Hal ini

dilakukan agar data mudah untuk diolah oleh algoritma *machine learning*. Normalisasi ini adalah mengubah skala yang pada dataset agar lebih seragam atau mendekati distribusi normal. Dengan mengubah skala antara 1-0.

Dari penjelasan tersebut dapat disumpulkan tahapan-tahapan melalui *flowchart* berikut ini.

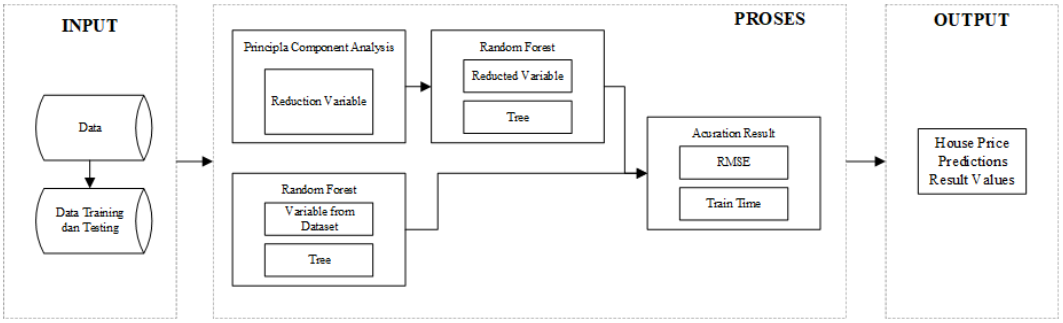


Gambar 3. 3 *Flowchart* Alur Penelitian dengan PCA dan *Random Forest*

3.2 *Desain System*

Desain sistem ini menggambarkan bagaimana sistem yang akan dibangun bekerja. Dalam kasus ini input yang digunakan adalah spesifikasi properti, dimana didalam spesifikasi properti tersebut memiliki beberapa variabel fitur yang dimiliki oleh properti dan harga dari properti tersebut. Metode yang digunakan adalah

metode *random forest*. Kemudian untuk optimasi dalam penelitian ini digunakan *Principal component analysis*. Metode ini cukup bagus digunakan untuk melakukan prediksi harga rumah berdasarkan fitur fiturnya. Kemudian output dari sistem yang akan dibangun adalah prediksi harga rumah.



Gambar 3. 4 Desain System

3.3 Experiment

3.3.1 *Principal Component Analysis*

Principe Component Analysis atau PCA adalah sebuah teknik analisis data yang digunakan untuk melakukan pengurangan dimensi pada sebuah dataset dengan tetap mempertahankan informasi pada dataset. PCA bekerja dengan mencari vektor utama (*principal components*) dari dataset yang memiliki variansi terbesar, yang kemudian digunakan untuk mewakili data. PCA umumnya digunakan ketika variabel dalam data memiliki kemiripan atau korelasi yang tinggi antar column. Kemiripan ini sering disebutnya data yang berulang atau redundant.

Berikut adalah langkah-langkah umum dalam melakukan PCA:

1. Normalisasi data: Data harus dalam bentuk skala yang sama agar perbandingan antar fitur tidak terpengaruh oleh skalanya.

2. Menghitung matrik kovariansi: Matrik kovariansi digunakan untuk menghitung variansi dan korelasi antar fitur dalam dataset.

   Varian (1 Atribut):

$$var(A_1) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})}{(n-1)} \quad (3.2)$$

   Kovarian (2 Atribut):

$$cov(A_1, A_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})}{(n-1)} \quad (3.3)$$

   Setelah diketahui varians dan kovariannya langkah selanjutnya yaitu menghitung matrik kovariannya. Berikut adalah rumusnya

$$C^{nxn} = \left(C_{i,j}\right), Where\ C_{i,j} = cov(A_1, A_2) \quad (3.4)$$

3. Menghitung *eigenvectors* dan *eigenvalue*s: *Eigenvectors* adalah vektor utama yang memiliki variansi terbesar dan *eigenvalue*s adalah variansi yang terkait dengan masing-masing *eigenvector*. *Eigenvectors* dan *eigenvalue*s dari matrik kovariansi C dapat ditemukan dengan memecahkan persamaan *eigenvalue* berikut:

$$\mathrm{Cv} = \lambda \mathrm{v} \quad (3.5)$$

   di mana λ adalah *eigenvalue* dan v adalah *eigenvector*

4. Pemilihan *eigenvectors*: *Eigenvectors* dengan variansi terbesar dipilih sebagai principal components.
5. Proyeksi data: Data dapat dikompresi dengan memproyeksikan data ke dalam subruang spasial yang didefinisikan oleh principal components.

3.3.2 *Random Forest*

Algoritma *random forest* adalah salah satu algortima yang termasuk dalam supervised learning. Algoritma ini bisa menyelesaikan masalah klasifikasi ataupun regresi. Selain sederhana algoritma ini mumpuni untuk digunakan dalam menyelesaikan masalan klasifikasi ataupun regresi.

Algoritma *Random Forest* termasuk dalam *ensemble learning*. *Ensemble learning* ini bekerja dengan beberapa metode yang bekerja sama untuk melakukan kinerjanya. Karena bekerja dengan bersama kinerja dari model ini bisa dikatakan baik dari pada model yang bekerja dengan sendirinya. Karena model ini bekerja dengan beberapa metode yang bekerja sama hasil akhir dari nilainya akan digabung. Untuk kasus regresi penggabungan nilai menggunakan nilai mean atau rata-rata, untuk kasus klasifikasi menggunakan pemunculan nilai mode yang terbanyak.

Ada 2 pendekatan yang sering ditemui dalam *ensemble learning* yaitu *boosting* dan *bagging. Boosting* atau *bagging* adalah teknik yang digunakan untuk memilih data pelatihan secara random. Sehingga dengan teknik ini sampel dataset yang digunakan dalam melatih model berbeda dan satu sampel dengan sampel yang lain berbeda. Dan pastinya model satu dengan model yang lain memiliki hasil yang berbeda.

Pada langkah selanjutnya, sampel acak sebanyak 2/3 dari data diambil, dari mana data tersebut kemudian digunakan sebagai pohon. Kemudian buat pohon dari beberapa pohon ini dari setiap sampel dan bentuk hutan. Dari data random forest, setiap variabel dihitung untuk mendapatkan nilai variabel penting (VI) atau variabel yang paling berpengaruh. VI ini dipengaruhi oleh nilai kesalahan *out-of-bag* (OOB)

yang dihasilkan dari data uji yang diuji. Penjelasan pola kesalahan OOB adalah sebagai berikut

$$errOOB = \frac{1}{n-z}\sum_{i=1}^{n-z}(y_i - \hat{y}_i)^2 \quad (3.6)$$

Diketahui,

n = data observasi
z = *sample* (data testing) untuk membentuk tree
y_i = data kei-*i* (data testing)
$\hat{y}_i$ = prediksi ke-*i*.

Kemudian untuk persamaan untuk mendapatkan variable importan (VI) diperoleh dari rumus sebagai berikut

$$VI\left(x^j\right) = \frac{1}{2}\sum_{t}^{s}(errOOB_t^j - errOOB_i) \quad (3.7)$$

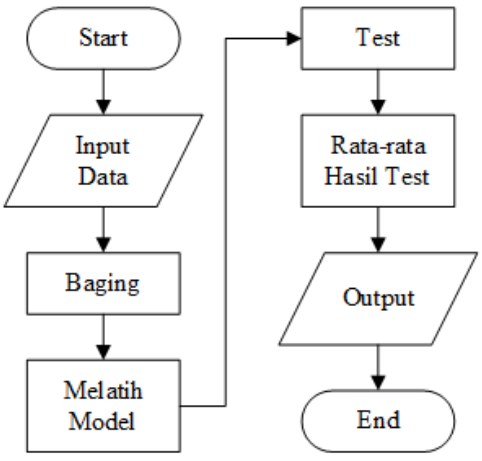
Dimana

$VI\left(x^j\right)$ = Variabel important variabel ke-j

S = jumah tree

$errOOB_t^j$ = error pada tree ke-t dengan variabel j dan t

$errOOB_i$ = error pada tree ke-t variabel ke-t



Gambar 3. 5 Alur Algoritma *Random Forest*

Gambar tersebut menjelaskan alur dari bagaimana algoritma *random forest* bekerja. Proses dari algoritma *random forest* dijelaskan sebagai berikut

1. Input Data yang sudah masuk kemudian di split menjadi data *training* dan data *testing*.
2. Kemudian data *training* masuk ke tahap bagging atau *bootstrap aggregating*, untuk membuat sampel data latih yang berbeda dengan sampel yang lain.
3. Kemudian model yang digunakan untuk melatih setiap sampel menggunakan algoritma decission tree, pada dasarnya *random forest* ini adalah gabungan dari beberapa algoritma *decission tree*.
4. Kemudian model yng sudah dilatih akan diuji dengan data uji, karena sistem dari *ensemble learning* ini adalah bekerja dengan beberapa metode yang bekerja sama untuk melakukan kinerjanya maka hasil dari setiap model akan dirata-rata untuk kasus regresi, dan untuk kasus klasifikasi yaitu menggunakan pemunculan nilai mode yang terbanyak.

3.4 Evaluasi

Setelah tahap penerapan model, langkah selanjutnya yaitu melakukan evaluasi untuk melakukan uji model. Untuk memastikan metode yang digunakan baik atau tidak yaitu dengan mengujinya. Untuk mengevaluasi model regresi secara teknis hanya menghitung selisih antara nilai sebenarnya dan nilai prediksi dalam hal ini bisa disebut error.

Namun sebelum melangkah pada evaluasi mengecek nilai error pada model yang sudah dilatih perlu dilakukan scaling terhadap data numerik pada dataset

testing. Hal ini dilakukan agar menghindari kebocoran data. Hal ini harus dilakukan agar skala antara data latih dan data uji sama dan kita bisa melakukan evaluasi.

Untuk menguji, matrik yang akan digunakan yaitu matrik RMSE, berikut adalah persamaanya

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(At - Ft)^2}{n}} \tag{3.8}$$

Dimana:

At = Nilai aktual/ nilai sebenarnya

Ft = Nilai prediksi

N = Jumlah dataset

3.5 Research Instrument

Bagian ini adalah bagian yang penting dalam sebuah penelitian. Baik atau tidaknya hasil dari uji coba yang dilakukan tergantung pada instrumen penelitian. Dalam penelitian ini variabel bebas adalah data penjualan rumah yang terdiri dari beberapa variabel yaitu id, jumlah kamar tidur, jumlah kamar mandi, luas tanah, harga per meter, alamat, luas bangunan, sertifikat, interior, parkir, listrik, dan harga. Data yang digunakan diperoleh dari hasil *scraping* data penjualan rumah dari website rumah.com. Untuk variabel terikat adalah kinerja dari hasil penelitian yaitu tingkar error. Sedangkan untuk variabel antara dalam penelitian ini yaitu hasil dari prediksi harga rumah.

BAB IV
PEMBAHASAN

Pada bagian ini dipaparkan hasil-hasil penelitian sekaligus pembahasan yang komprehensif. Penjelasan langkah langkah dan hasil analysisnya. Berikut adalah penjelasannya

4.1 *Business Understanding*

Bisnis dibidang properti mempunyai keunggulan yang membuat setiap orang yang terjun didalamnya perlu hati hati. Terutama pada orang yang ingin berinvestasi di bidang properti. Memang investasi menjadi hal yang menarik, khususnya investasi di bidang properti. Sudah sejak tahun 2011 investasi properti ini meningkat secara signifikan baik secara on demand maupun secara penjualan (R.M.A. van der Schaar, 2015).

Pihak pengembang juga harus berhati hati dalam menentukan harga properti. Perlu diketahui setiap tahunnya baik jangka pendek ataupun jangka panjang harga properti semakin naik dan bahkan hampir tidak pernah turun (Feng & Jones, 2015). Salah memperhitungkan harga properti membuat bisnis dan investasi bisa menjadi rugi. Resiko ini sangatlah tidak bagus jika dialami oleh setiap pebisnis. Tentu saja setiap bisnis selalu mengejar yang namanya profit. Penting bagi seorang pebisnis, terutama dibidang properti untuk mengetahui dan mampu untuk memprediksi harga properti. Sehingga pebisnis dapat mendapatkan profit sebesar mungkin.

4.1.1 *Problem Statements and Goals*

Dari pemaparan tersebut dapat, perlu membuat sebuah sistem prediksi harga rumah dan berikut adalah permasalahan yang akan diselesaikan

1. Dari fitur yang ada, fitur apa yang paling berpengaruh terhadap penjualan harga rumah.
2. Berapa harga pasar rumah dengan kondisi karakteristik terterntu?

Untuk menjawab pertanyaan tersebut tujuan dari masalah tersebut yaitu sebagai berikut

1. Mengetahui fitur yang paling berpengaruh terhadap penjualan rumah
2. Mengetahui harga pasar rumah berdasarkan kondisi karakteristik tertentu.

4.1.2 Metodologi

Dari pemaparan metode penelitian di BAB 3, tujuan awal yaitu untuk melakukan prediksi harga rumah. Variabel harga merupakan variabale yang bersifat kontinu. Dalam kasus variabel yang bersifat kontinu, permasalahan yang akan diselesaikan adalah permasalah regresi. Oleh karena itu metodologi pada penelitian ini adalah membangun model regresi dengan targat harga rumah.

4.1.3 Matrik

Salah satu cara untuk mengetahui apakah model yang digunakan baik atau tidak yaitu dengan melakukan testing. Untuk kasus regresi biasanya matrik yang digunakan untuk melakukan testing yaitu menggunakan *error*. Untuk melakukan testing *error* bisa menggunakan *root mean square error (RMSE).* Matrik ini akan mengukur seberapa jauh hasil prediksi dengan hasil sebenarnya.

4.2 *Analytic Approach*

Untuk melakukan prediksi harga rumah, metode yang akan digunakan yaitu metode *random forest*. Ja'afar *et al.*, (2021) Melakukan penelitian *literature review* berkaitan dengan prediksi harga rumah. Ditemukan bahwa metode terbaik yang

ditemukan adalah metode *random forest*. *Random forest* mempunyai performa yang bagus dari pada yang lain dalam kontek untuk prediksi rumah.

Metode *random forest* ini memiliki kemampuan yang bagus untuk melakukan prediksi harga rumah. *Random forest* tergolong konsep pembelajaran *ensemble learning*, yaitu konsep yang merata-ratakan hasil dari beberapa pohon keputusan/*decission tree* yang diterapkan dari kumpulan data untuk meningkatkan akurasi. Dengan membuat jumlah pohon keputusan yang besar, metode ini dapat memberikan tingkat akurasi yang tinggi dan bisa menghindari masalah *overfitting*, namun hal itu mempunyai kelemahan yaitu meningkatkan waktu pelatihan yang lama (Adetunji et al., 2022). Selain itu pengacakan pada sampel *bagging* dan pemilihan fitur pada *random forest* cenderung memilih fitur yang tidak informatif untuk pemisahan node (Nguyen et al., 2015). Ini membuat *random forest* memiliki akurasi yang buruk saat bekerja dengan data dimensi tinggi.

Salah satu cara untuk mengurangi kekurangan tersebut yaitu dengan memilih fitur yang informatif, karena Pemilihan fitur merupakan langkah penting untuk mendapatkan performa yang baik untuk model (Nguyen et al., 2015). Selain itu pengurangan jumlah fitur dalam metode *random forest* dapat mempercepat kinerja dari model *random forest* tersebut. Salah satu cara yang digunakan untuk mengurangi fitur tanpa harus menghapus fitur yang lain yaitu dengan menggabungkan beberapa fitur yaitu menggunakan *Principal Component Analysis* (Gardner & Lo, 2021).

4.3 *Data Requirement*

Diberbagai negara *House Price Index* (HPI) sering digunakan untuk menghitung kenaikan harga rumah (Garriga et al., 2021), namun kondisi fisik, konsep, lokasi juga mempengaruhi dalam menentukan harga rumah (Nur et al., 2017). Karakteristik fisik yang dimiliki rumah seperti jumlah kamar tidur, kamar mandi, ukuran bangunan dapat mempengaruhi harga rumah (Kang et al., 2021).

Maka dari itu untuk menentukan harga rumah dalam penelitian ini yaitu menggunakan kondisi fisik dari suatu rumah.

4.4 *Data Collection*

Pengumpulan data pada penelitian ini menggunakan data publik, diambil dari hasil *scraping* dari website penjualan rumah kota malang yaitu dari website rumah.com yang dilakukan pada tanggal 17 februari 2023. Didalam dataset tersebut terdapat 6130 baris dan 11 kolom. Bentuk file yang tersedia berbentuk file xlsx.

4.5 *Data Understanding*

4.5.1 *Exploratory Data Analysis* - Deskripsi Variabel

Berikut adalah deskripsi variabel dari data yang akan digunakan untuk melakukan prediksi harga rumah.

Tabel 4. 1 Dataset dan Diskripsi

No	Variabel	Type	Tipe Data	Deskripsi
1	jumlah kamar tidur	Fitur Rumah	float	Jumlah kamar tidur di rumah untuk dijual
2	jumlah kamar mandi	Fitur Rumah	float	Luas rumah dalam meter persegi
3	luas tanah	Fitur Rumah	float	Luas rumah dalam meter persegi
4	harga per meter	Fitur Rumah	float	Nilai harga sebuah rumah dihitung per meter persegi

5	alamat	Fitur Rumah	Object	Lokasi rumah yang dijual
6	luas bangunan	Fitur Rumah	float	Luas bangunan dalam meter persegi
7	Sertifikat	Fitur Rumah	Object	Sertifikat penjualan rumah
8	Interior	Fitur Rumah	Object	Jenis interior yang ada di dalam rumah
9	parkir	Fitur Rumah	float	Jumlah tempat parkir di rumah
10	listrik	Fitur Rumah	float	Tegangan listrik di rumah dijual dalam satuan watt
11	harga	Target	float	Nilai harga rumah yang dijual

4.5.2 *Exploratory Data Analysis* – Menangani *outlier*

Tahap ini akan dilihat data mana yang memiliki nilai *missing value*. Berikut adalah hasil deteksi *missing value*

```
kamar tidur            0
kamar mandi            0
luas tanah(m2)         0
harga per m            0
Alamat                 0
luas Bangunan(m2)      0
sertifikat             0
interior            2599
parkir              2104
listrik              909
harga                  0
dtype: int64
```

Gambar 4.1 *Missing Value*

Dari gambar 4.1 dapat dilihat bahwa data listrik, parkir dan interior memiliki nilai *missing value*, maka dari itu baris yang memiliki nilai *missing value* tersebut akan dihapus.

4.5.3 *Exploratory Data Analysis* – Menangani *Outlier*

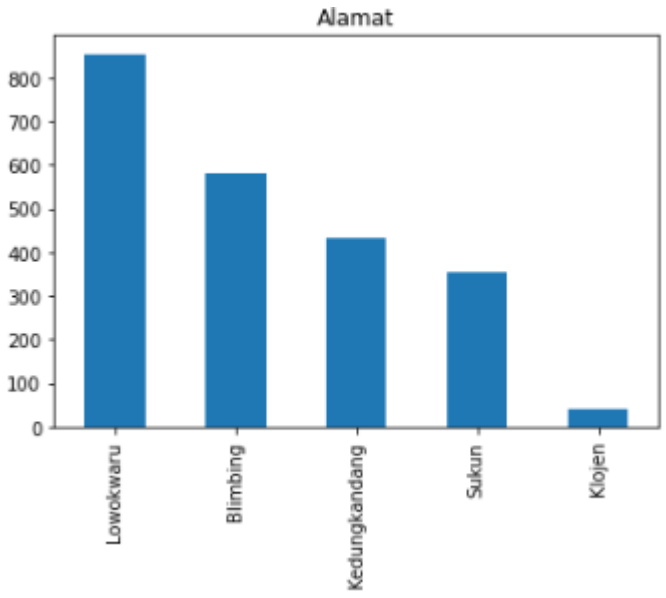
Tahap ini adalah tahap dimana mendeteksi *outlier*. Kemudian setelah dicek apakah ada *outlier* atau tidak, untuk menangani *outlier* disini menggunakan IQR

method, setelah fungsi dijalankan baris yang memiliki nilai *outlier* dan *outlier* ini akan dihapus.

IQR method (*Interquartile Range method*) adalah salah satu teknik untuk menentukan apakah suatu nilai adalah *outlier* atau tidak. IQR didasarkan pada konsep bahwa nilai yang berada diluar rentang interquartil (Q3-Q1) dapat dikategorikan sebagai *outlier*.

4.5.4 *Exploratory Data Analysis – Univariate Analysis*

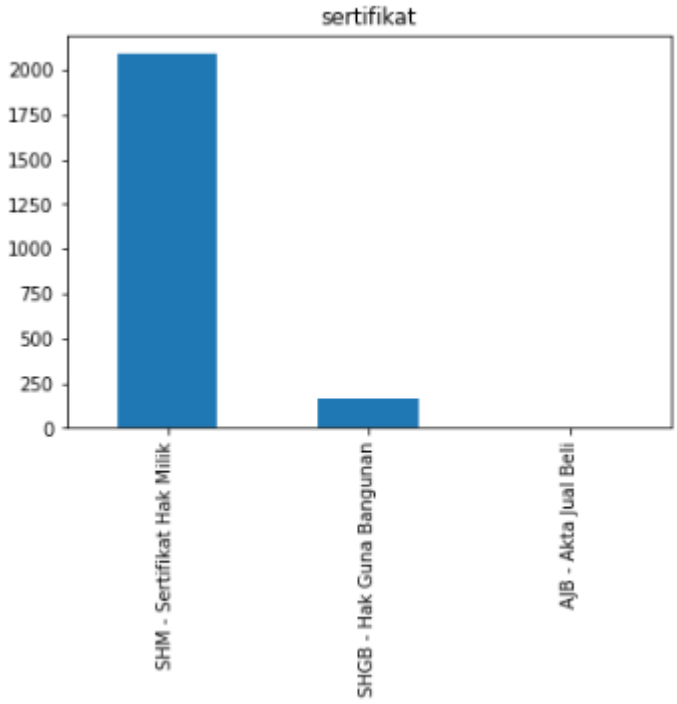
Tahap ini adalah tahap untuk menganalisis dataset yang digunakan. Dalam dataset yang digunakan ada 3 jenis data yang bersifat kategorik yang masih digunakan, yaitu data alamat, interior, sertifikat. Untuk fitur numerik akan dianalisis fitur target yaitu variabel harga



Gambar 4. 2Variabel Alamat

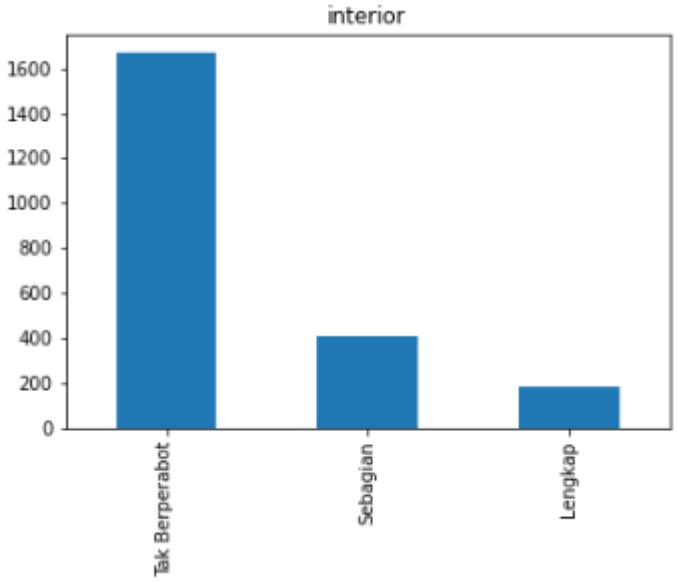
Dari visualisasi di gambar 4.2 tersebut dapat disimpulkan bahwa Kecamatan Lowokwaru memiliki jumlah penjualan rumah terbanyak yaitu sejumlah 854 atau

sebesar 37.8% dan kecamatan Klojen memiliki jumlah penjualan rumah yang paling sedikit yaitu sejumlah 41 atau sebesar 1.8%.



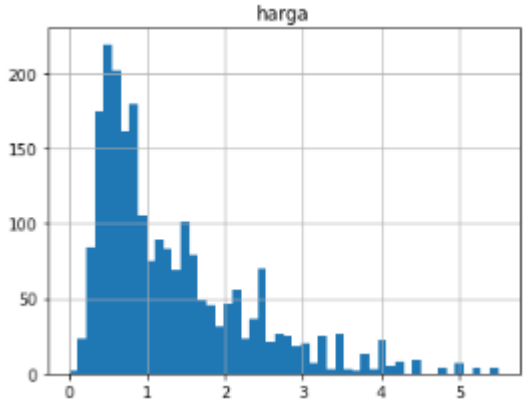
Gambar 4. 3 Variabel Sertifikat

Dari visualisasi di gambar 4.3 tersebut dapat disimpulkan bahwa rumah yang mempunyai sertifikat SHM – Sertifikat Hak Milik memiliki jumlah terbanyak yaitu dengan jumlah 2088 atau sebesar 92.5% dan AJB – Akta Jual Beli memiliki jumlah terkecil yaitu sebanyak 8 atau sebesar 0.4%.



Gambar 4. 4 Variabel Interior

Dari visualisasi di gambar 4.4 tersebut dapat disimpulkan bahwa rumah yang mempunyai interior tidak berperabot memiliki jumlah terbanyak yaitu dengan jumlah 1666 atau sebesar 73.8% dan interior lengkap memiliki jumlah terkecil yaitu sebanyak 184 atau sebesar 8.1%.



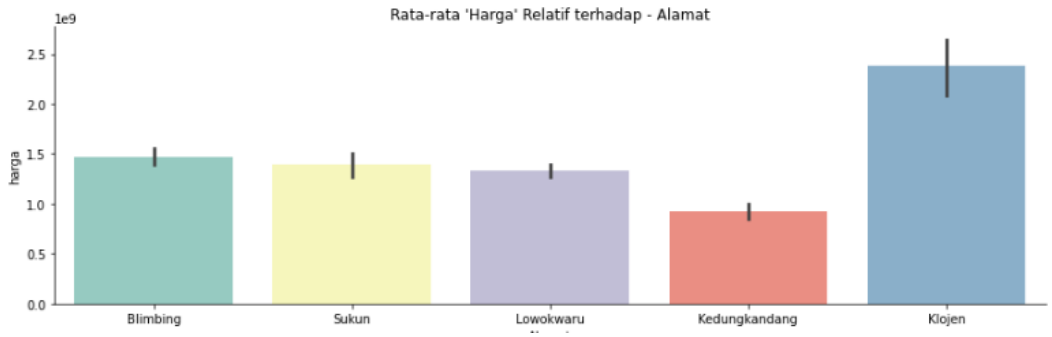
Gambar 4. 5 Variabel Harga

Dari visualisasi gambar 4.5 tersebut dapat diperoleh beberapa informasi, yaitu

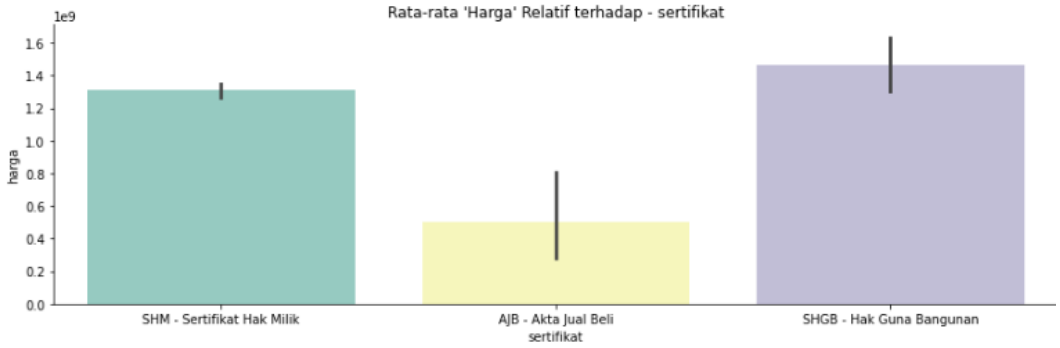
- Peningkatan harga sebanding dengan penurunan jumlah sampel/ jumlah rumah. Hal ini dapat dilihat dari grafik mengalami penurunan seiring dengan banyaknya jumlah sampel
- Distribusi harga miring ke kanan, ini menandakan bahwa hal ini akan berpengaruh terhadap model.

4.5.5 *Exploratory Data Analysis – Multivariate Analysis*

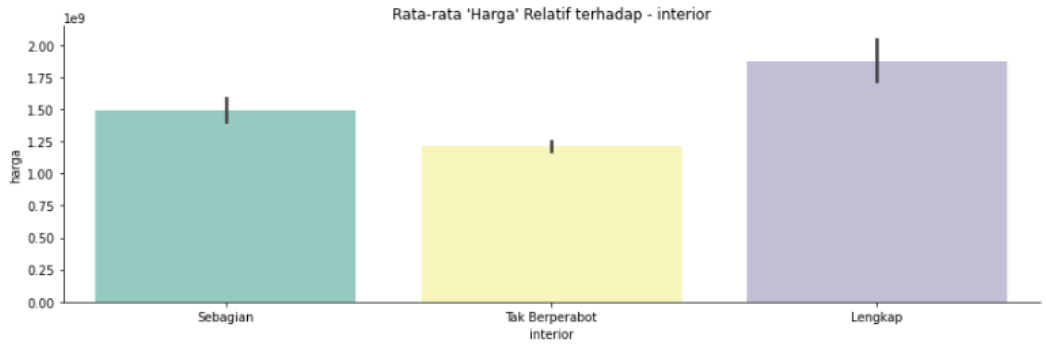
Tahap ini tahap untuk mengetahui hubungan antara dua atau lebih variabel yang digunakan pada data.



Gambar 4. 6 Rata-rata ‘Harga’ Relatif terhadap – Alamat



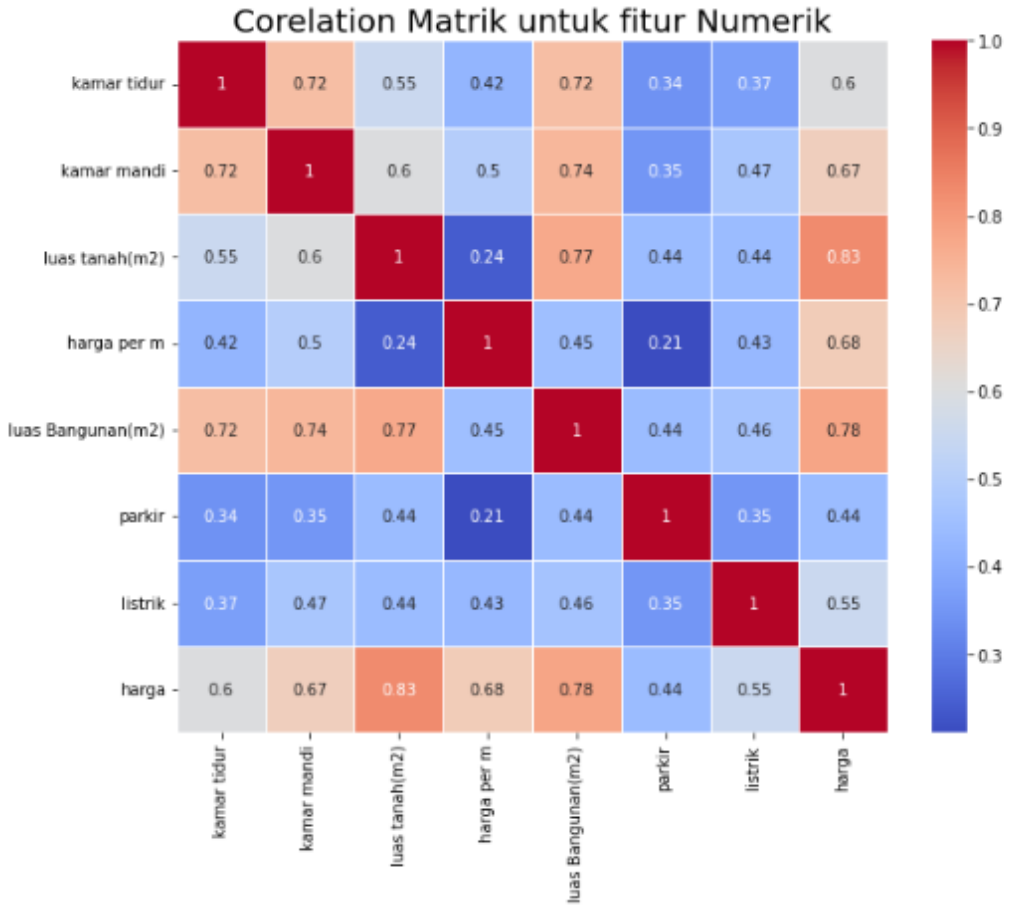
Gambar 4. 7 Rata-rata ‘Harga’ Relatif terhadap - Sertifikat



Gambar 4. 8 Rata-rata ‘Harga’ Relatif terhadap – interior

Dari visualiasi gambar 4.6, 4.7 dan 4.8 tersebut diperoleh beberapa informasi yaitu:

- Pada variabel alamat, rata-rata harga cenderung mirip antara kecamatan Sukun, Blimbing dan Lowokwaru, berkisar diangka 1.5 Miliar. Sedangkan yang beralamat di Koljen memiliki rata-rata harga cenderung tinggi berkisar 2.5 Miliar
- Pada variabel sertifikat, rata-rata harga terhadap rumah yang memiliki sertifikat SHGB dan SHM cenderung tinggi daripada rumah yang memiliki sertifikat AJB.
- Pada variabel interior, rata-rata harga terhadap rumah yang memiliki interior lengkap cenderung tinggi daripada yang tidak berperabot atau sebagian.



Gambar 4. 9 Corelation Matrik Fitur Numerik

Matriks korelasi adalah ukuran statistik yang menggambarkan hubungan antara dua variabel atau lebih dalam bentuk matriks. Matriks korelasi dapat memberikan gambaran seberapa kuat atau lemahnya hubungan antara variabel-variabel tersebut. Berikut adalah langkah langkah umum yang digunakan untuk menentukan korelasi matrik.

1. Menentukan variabel yang akan dianalisis, dalam hal ini semua variabel numerik akan dianalisis.
2. Menghitung nilai korelasi antara setiap pasangan variabel. Korelasi biasanya dihitung menggunakan koefisien korelasi Pearson, yang mengukur hubungan linier antar variabel. Koefisien korelasi Pearson

memiliki rentang nilai dari -1 hingga 1. Nilai positif menunjukkan hubungan positif, nilai negatif menunjukkan hubungan negatif, dan nilai 0 menunjukkan tidak ada hubungan linier.

3. Menghitung koefisien korelasi antara setiap pasangan variabel menggunakan rumus korelasi Pearson. Berikut adalah rumusnya

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x^2))\,(n\sum y^2 - (\sum y^2))}} \qquad (4.\ 1)$$

Dimana

r = Koefisien Pearson
n = jumlah pasangan observasi
∑ xy = jumlah produk dari observasi berpasangan
∑ x = jumlah skor x
∑ y = jumlah skor y
∑ x2 = jumlah skor x kuadrat
∑ y2 = jumlah skor y kuadrat

4. Masukkan koefisien korelasi ke dalam matriks korelasi. Matriks korelasi memiliki dimensi yang sama dengan jumlah variabel yang akan dianalisis. Diagonal matriks mengandung 1 karena koefisien korelasi antara variabel dan sin adalah 1
5. Menafsirkan matriks korelasi. Nilai koefisien korelasi yang mendekati 1 menunjukkan adanya hubungan positif yang kuat antara kedua variabel. Nilai yang mendekati -1 menunjukkan hubungan negatif yang kuat antara dua variabel. Nilai yang mendekati nol menunjukkan bahwa tidak ada hubungan linier antara kedua variabel tersebut.

Dari visualisasi gambar 4.9 tersebut dapat disimpulkan bahwa variabel atau fitur luas tanah dan luas bangunan memiliki pengaruh yang tinggi terhadap harga, sedangkan variabel parkir memiliki korelasi yang rendah terhadap harga.

4.6 *Data Preparation*

4.6.1 *Data Preprocessing – One Hot Encoding*

Pada tahap ini adalah tahap dimana data yang bersifat kategorik akan dirubah menjadi data numerik. Disini memiliki 3 data kategorik yaitu alamat, interior dan sertifikat. Tahap ini nantinya akan dibuatkan variabel baru berdasarkan banyaknya jenis data kategorik. Didalam variabel alamat terdapat 5 jenis, pada variabel interior terdapat 3 jenis dan pada variabel sertifikat terdapat 3 jenis, jadi nanti terdapat 11 variabel baru yang isinya nilai 0 dan 1 berdasarkan nilai variabel tersebut.

Tabel 4. 2 Dataset Setelah Proses *One Hot Encoding*

Variable	Type	Data Type
Number of Bedrooms	Features	Float
Total Bathrooms	Features	Float
Surface Area	Features	Float
Price per Meter	Features	Float
Cut_sukun	Features	Float
Cut_kedungkandang	Features	Float
Cut_klojen	Features	Float
Cut_lowokwaru	Features	Float
Cut_Blimbing	Features	Float
Building Area	Features	Float
Cut_SHGB-Hak Guna Bangunan	Features	Float
Cut_AJB – Akta Jual Beli	Features	Float
Cut_SHM – Sertifikat Hak Milik	Features	Float
Cut_Tak Berperabot	Features	Float
Cut_Lengkap	Features	Float
Cut_Sebagian	Features	Float
Electricity	Features	Float
price	Target	Float

4.6.2 *Data Preprocessing* – Reduksi Dimensi dengan PCA

Teknik pengurangan dimensi adalah prosedur yang mengurangi jumlah fitur sambil mempertahankan informasi dalam data. Teknik pengurangan dimensi yang paling populer adalah analisis komponen utama, atau PCA. Ini adalah teknik untuk mengurangi ukuran, mengekstrak fitur, dan mengubah data dari "ruang n-dimensi" menjadi sistem koordinat baru dengan dimensi m, di mana m kurang dari n.

PCA bekerja dengan metode aljabar linier. Himpunan data diasumsikan paling penting (primer) dengan arah varian terbesar. PCA biasanya digunakan ketika variabel dalam data sangat berkorelasi. Korelasi yang tinggi ini menunjukkan bahwa data tersebut berulang. Oleh karena itu, teknik PCA digunakan untuk mereduksi variabel asli menjadi sejumlah kecil variabel baru yang tidak berkorelasi linier, yang disebut komponen utama (PC). Komponen utama ini dapat menangkap sebagian besar varian dari variabel asli. Sehingga ketika teknik PCA diterapkan pada data, hanya menggunakan komponen utama dan mengabaikan yang lainnya.

Di bawah ini adalah penjelasan dari masing-masing komponen utama (PC):

- komponen utama pertama mewakili arah varians terbesar dalam data. Ini mengumpulkan informasi paling banyak dari semua karakteristik data.
- komponen utama kedua menangkap sebagian besar data yang tersisa setelah komponen utama pertama.
- komponen utama ketiga mengumpulkan sebagian besar data yang ditinggalkan oleh komponen utama pertama, komputer kedua, dst.

Dari proses *one hot encoding*, variabel yang terbuat menjadi semakin banyak, hal ini membuat proses pelatihan semakin lama dan kurang efektif. Terdapat 11 variabel baru yaitu Cut_Blimbing, Cut_kedungkandang, Cut_klojen, Cut_lowokwaru, Cut_sukun yang memiliki informasi yang sama yaitu alamat, sedangkan Cut_SHGB-Hak Guna Bangunan, Cut_AJB – Akta Jual Beli, Cut_SHM – Sertifikat Hak Milik memiliki informasi yang sama yaitu sertifikat, sedangkan Cut_lengkap, Cut_Sebagian, Cut_Tak Berperabot memiliki informasi yang sama yaitu interior. PCA akan mereduksi dimensi, ke dalam sistem berkoordinat baru. Dalam kasus ini variabel baru akan direduksi berdasarkan variabel aslinya, yaitu alamat, interior dan sertifikat. Berikut adalah data sebelum masuk ke tahap PCA. Untuk contoh keseluruhan data terdapat dilampiran.

Tabel 4. 3 Sample Data Sebelum PCA

cut_Blimbing	cut_Kedungkandang	cut_Klojen	cut_Lowokwaru	cut_Sukun	cut_AJB	cut_SHGB	cut_SHM	cut_Lengkap	cut_Sebagian	cut_Tak Berperabot
1	0	0	0	0	0	0	1	0	1	0
0	0	0	1	0	0	0	1	0	1	0
0	0	0	1	0	0	0	1	0	0	1
0	0	0	1	0	0	0	1	0	0	1
0	1	0	0	0	0	0	1	0	1	0
1	0	0	0	0	0	0	1	0	0	1
0	1	0	0	0	0	0	1	0	0	1
0	1	0	0	0	0	0	1	0	0	1
0	0	0	1	0	1	0	0	0	0	1
1	0	0	0	0	0	0	1	0	1	0
0	1	0	0	0	0	0	1	1	0	0
0	0	0	1	0	0	0	1	1	0	0
0	0	0	1	0	0	0	1	1	0	0
0	0	0	0	1	0	0	1	1	0	0
0	0	1	0	0	0	0	1	0	0	1

0	0	1	0	0	0	0	1	0	0	1
1	0	0	0	0	1	0	0	0	0	1
0	0	0	1	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	0	0	1
0	0	0	1	0	0	1	0	0	0	1
0	0	0	0	1	0	1	0	0	0	1
0	0	0	0	1	0	1	0	0	0	1
0	0	0	1	0	0	1	0	0	0	1

Dalam kasus ini proses PCA akan dibagi menjadi 3 bagian, yang pertama akan mereduksi terlebih dahulu untuk fitur alamat dan yang kedua untuk fitur sertifikat dan yang ketiga untuk fitur interior.

A. Fitur Alamat

1. Normalisasi Data

Tahap normalisasi ini digunakan untuk merubah data menjadi skala yang sama agar perbandingan antar fitur tidak terpengaruh oleh skalanya. Dalam kasus ini, yang terdapat pada fitur alamat sudah memiliki skala data yang sama. Berikut adalah contoh datanya

Tabel 4. 4 Contoh data normalisasi fitur alamat

cut_Blimbing	cut_Kedungkandang	cut_Klojen	cut_Lowokwaru	cut_Sukun
1	0	0	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	1	0	0	0
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	1	0

1	0	0	0	0
0	1	0	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	1	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	1	0
1	0	0	0	0
0	0	0	1	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	1
0	0	0	1	0

2. Menghitung Matrik Kovarian/Korelasi

Sebelum menghitung matrik kovarian, perlu diketahui matrik kovarian ini digunakan untuk menghitung variansi dan korelasi antar fitur dalam dataset. Nantinya nilai dari matrik varian kovarian berfungsi sebagai masukan untuk mendapatkankan nilai *eigen* dan *vektor eigen*.

Untuk menghitung matrik kovarian, terlebih dahulu untuk menghitung nilai varian dan nilai kovarian.

Varian (1 Atribut):

$$var(A_1) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})}{(n-1)} \qquad (4.\ 2)$$

Kovarian (2 Atribut):

$$cov(A_1, A_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})}{(n-1)} \qquad (4.\ 3)$$

Setelah diketahui varians dan kovariannya langkah selanjutnya yaitu menghitung matrik kovariannya. Berikut adalah rumusnya

$$C^{nxn} = (C_{i,j}), Where\ C_{i,j} = cov(A_1, A_2) \quad (4.\ 4)$$

Berikut untuk nilai varian dari masing masing variabel

```
array([3.20449173e-01, 2.19495270e-01, 1.68047804e-01, 2.22429467e-02,
       2.32677342e-31])
```

Gambar 4. 10 Nilai Varian dari masing masing variabel fitur alamat

Berikut untuk nilai matrik kovariannya

```
array([[ 0.19096969, -0.04905121, -0.00466612, -0.09719197, -0.04006039],
       [-0.04905121,  0.15451133, -0.00346741, -0.07222369, -0.02976901],
       [-0.00466612, -0.00346741,  0.01783586, -0.00687047, -0.00283186],
       [-0.09719197, -0.07222369, -0.00687047,  0.23527159, -0.05898547],
       [-0.04006039, -0.02976901, -0.00283186, -0.05898547,  0.13164673]])
```

Gambar 4. 11 Nilai Kovarian matrik fitur alamat

Nilai matrik kovarian tersebut digunakan menghitung *eigenvector* dan *eigenvalue*nya.

3. Menghitung *Eigenvector* dan *Eigenvalue*

Untuk menghitung *eigenvalue* dan eigenvektor diperlukan nilai dari kovarian matrik. Berikut adalah persamaan untuk menghitung *eigenvalue* dan *eigenvektor*.

$$\mathrm{Cv} = \lambda \mathrm{v} \quad (4.\ 5)$$

Turunan dari rumus tersebut bisa menjadi seperti ini. M tersebut adalah nilai matrik kovarian yang telah dihitung ditahap sebelumnya.

$$(M - \lambda I)\, v = 0 \qquad (4.6)$$

Setiap nilai *eigenvalue* atau " λ " harus memenuhi persamaan determinan. Setelah nilai *eigen value* ditemukan maka hasil dari *eigenvalue* dikembalikan dipersamaan $Cv = \lambda v$ untuk menghitung nilai *eigenvector*

$$|M - \lambda I| = 0 \qquad (4.7)$$

Dari persamaan tersebut "v" adalah *eigenvektor* dan " λ " adalah *eigenvalue*. Dari hasil perhitungan diperoleh nilai *eigenvalue* yang dihitung dari matriks kovariansi data yang telah diproses oleh PCA sebagai berikut

```
array([3.20449173e-01, 2.19495270e-01, 1.68047804e-01, 2.22429467e-02,
       2.32677342e-31])
```

Gambar 4. 12 matriks kovariansi fitur alamat

Untuk nilai *eigenvektor* sebagai berikut

```
array([[-0.51364018, -0.18709157, -0.00762386,  0.8286399 , -0.12028428],
       [-0.67390454,  0.65784316,  0.00886913, -0.23409121,  0.24128345],
       [-0.1836778 , -0.52973951,  0.00770395, -0.11434964,  0.820063  ],
       [-0.21974737, -0.22718974,  0.89431754, -0.21326366, -0.23411677],
       [ 0.4472136 ,  0.4472136 ,  0.4472136 ,  0.4472136 ,  0.4472136 ]])
```

Gambar 4. 13 Nilai *Eigenvektor* fitur alamat

4. Pemilihan *Eigenvector*

Setelah ketemu nilai *eigenvector* dan *eigenvalue*, langkah selanjutnya mengurutkan *eigenvector* dari *eigenvalue* yang paling besar ke yang paling kecil. Karena ada beberapa nilai *eigenvector* dari beberapa *eigenvalue*, maka perlu dihitung proporsi dari setiap *eigenvalue* mana yang nilai paling besar. *Eigenvalue* yang memiliki proporsi paling besar ini memiliki informasi yang

paling tinggi, yang nantinya nilai *eigenvector* dari *eigenvalue* yang memiliki proporsi paling tinggi akan dijadikan pengali untuk mereduksi data.

Setelah menerapkan class PCA, langkah selanjutnya yaitu mengetahui proporsi informasi dari komponen.

```
array([0.439, 0.301, 0.23 , 0.03 , 0.   ])
```

Gambar 4. 14 Proporsi Informasi Komponen fitur alamat

Arti dari output tersebut adalah 43,9 % informasi terdapat pada PC pertama, 30.1% terdapat pada PC kedua, 23% pada PC ketiga, 3% pada PC keempat dan 0% pada PC kelima.

5. Proyeksi Data

Pada hasil sebelumnya dapat dilihat pada PC pertama memiliki nilai tertinggi daripada yang lain. Maka dari itu maka akan mereduksi fitur dan hanya mempertahankan pada PC pertama saja. PC pertama ini akan menggantikan dari kelima fitur 'cut_Blimbing', 'cut_Kedungkandang', 'cut_Klojen', 'cut_Lowokwaru', 'cut_Sukun' menjadi fitur ‘alamat’. Berikut adalah hasil dari hasil reduksi

Tabel 4. 5 Contoh hasil dari nilai reduksi fitur alamat

Id	alamat
1	-0,6405
2	-0,24715
3	0,701776
4	0,701776
5	0,701776
6	0,701776
7	0,701776
8	0,701776
9	-0,24715
10	-0,24715

11	-0,24715
12	0,701776
13	-0,31396
14	-0,6405
15	0,701776
16	0,701776
17	-0,31396
18	-0,31396
19	-0,31396
20	-0,31396
21	0,701776
22	-0,6405
23	-0,31396
24	-0,6405
25	-0,6405

B. Fitur Sertifikat

1. Normalisasi Data

Tahap normalisasi ini digunakan untuk merubah data menjadi skala yang sama agar perbandingan antar fitur tidak terpengaruh oleh skalanya. Dalam kasus ini, yang terdapat pada fitur sertifikat sudah memiliki skala data yang sama. Berikut adalah contoh datanya

Tabel 4. 6 Contoh Data Hasil Normalisasi Fitur Sertifikat

cut_AJB	cut_SHGB	cut_SHM
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
1	0	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

0	0	1
0	0	1
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0

2. Menghitung Matrik Kovarian

Sebelum menghitung matrik kovarian, perlu diketahui matrik kovarian ini digunakan untuk menghitung variansi dan korelasi antar fitur dalam dataset. Nantinya nilai dari matrik varian kovarian berfungsi sebagai masukan untuk mendapatkankan nilai *eigen* dan *vektor eigen*.

Untuk menghitung matrik kovarian, terlebih dahulu untuk menghitung nilai varian dan nilai kovarian.

Varian (1 Atribut):

$$var(A_1) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})}{(n-1)} \quad (4.\ 8)$$

Kovarian (2 Atribut):

$$cov(A_1, A_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})}{(n-1)} \quad (4.\ 9)$$

Setelah diketahui varians dan kovariannya langkah selanjutnya yaitu menghitung matrik kovariannya. Berikut adalah rumusnya

$$C^{nxn} = (C_{i,j}), Where\ C_{i,j} = cov(A_1, A_2) \quad (4.\ 10)$$

Berikut untuk nilai varian dari masing masing variabel

```
array([1.34564580e-01, 5.24492078e-03, 1.03474094e-34])
```

Gambar 4. 15 Nilai Varian Fitur Sertifikat

Berikut adalah nilai dari matrik kovarian yang nantinya akan digunakan untuk menghitung *eigenvalue* dan *eigenvector*

```
array([[ 0.00353197, -0.0002543 , -0.00327767],
       [-0.0002543 ,  0.06662708, -0.06637278],
       [-0.00327767, -0.06637278,  0.06965045]])
```

Gambar 4. 16 Matrik Kovarian Fitur Sertifikat

3. Menghitung *Eigenvector* dan *Eigenvalue*

Untuk menghitung *eigenvalue* dan eigenvektor diperlukan nilai dari kovarian matrik. Berikut adalah persamaan untuk menghitung *eigenvalue* dan *eigenvektor*.

$$\mathrm{Cv} = \lambda \mathrm{v} \quad (4.\ 11)$$

Turunan dari rumus tersebut bisa menjadi seperti ini. M tersebut adalah nilai matrik kovarian yang telah dihitung ditahap sebelumnya.

$$(\mathrm{M} - \lambda \mathrm{I})\ \mathrm{v} = 0 \quad (4.\ 12)$$

Setiap nilai *eigenvalue* atau "λ" harus memenuhi persamaan determinan. Setelah nilai *eigen value* ditemukan maka hasil dari *eigenvalue* dikembalikan dipersamaan $\mathrm{Cv} = \lambda \mathrm{v}$ untuk menghitung nilai *eigenvector*

$$|\mathrm{M} - \lambda \mathrm{I}| = 0 \quad (4.\ 13)$$

Dari persamaan tersebut “v” adalah *eigenvector* dan “λ” adalah *eigenvalue*. Dari hasil perhitungan diperoleh nilai *eigenvalue* yang dihitung dari matriks kovariansi data yang telah diproses oleh PCA sebagai berikut

```
array([1.34564580e-01, 5.24492078e-03, 1.03474094e-34])
```

Gambar 4. 17 Matrik Kovarian Fitur Sertifikat

Untuk nilai *eigenvektor* sebagai berikut

```
array([[ 0.01653485,  0.69869435, -0.7152292 ],
       [ 0.81632914, -0.42248417, -0.39384497],
       [ 0.57735027,  0.57735027,  0.57735027]])
```

Gambar 4. 18 Nilai *Eigenvektor* Fitur Sertifikat

4. Pemilihan *Eigenvector*

Setelah ketemu nilai *eigenvector* dan *eigenvalue*, langkah selanjutnya mengurutkan *eigenvector* dari *eigenvalue* yang paling besar ke yang paling kecil. Karena ada beberapa nilai *eigenvector* dari beberapa *eigenvalue*, maka perlu dihitung proporsi dari setiap *eigenvalue* mana yang nilai paling besar. *Eigenvalue* yang memiliki proporsi paling besar ini memiliki informasi yang paling tinggi, yang nantinya nilai *eigenvector* dari *eigenvalue* yang memiliki proporsi paling tinggi akan dijadikan pengali untuk mereduksi data.

Setelah menerapkan class PCA, langkah selanjutnya yaitu mengetahui proporsi informasi dari komponen.

```
array([0.962, 0.038, 0.   ])
```

Gambar 4. 19 Proporsi Informasi Komponen Fitur Sertifikat

Arti dari output tersebut adalah 96,2% informasi terdapat pada PC pertama, 3,8% terdapat pada PC kedua, 0% pada PC ketiga.

5. Proyeksi Data

Pada hasil sebelumnya dapat dilihat pada PC pertama memiliki nilai tertinggi daripada yang lain. Maka dari itu maka akan mereduksi fitur dan hanya mempertahankan pada PC pertama saja. PC pertama ini akan menggantikan dari ketiga fitur 'cut_AJB - Akta Jual Beli','cut_SHGB - Hak Guna Bangunan','cut_SHM - Sertifikat Hak Milik' menjadi fitur ‘Sertifikat’. Berikut adalah hasil dari hasil reduksi

Tabel 4. 7 Contoh Hasil Reduksi Fitur Sertifikat

Id	sertifikat
1	-0,10403
2	-0,10403
3	-0,10403
4	-0,10403
5	-0,10403
6	-0,10403
7	-0,10403
8	-0,10403
9	-0,10403
10	-0,10403
11	-0,10403
12	-0,10403
14	-0,10403
15	-0,10403
16	-0,10403
17	-0,10403
18	-0,10403
19	-0,10403
20	-0,10403
21	-0,10403
22	0,62773
23	-0,10403
24	-0,10403
25	-0,10403

C. Fitur Interior

1. Normalisasi Data

Tahap normalisasi ini digunakan untuk merubah data menjadi skala yang sama agar perbandingan antar fitur tidak terpengaruh oleh skalanya. Dalam kasus ini, yang terdapat pada fitur interior sudah memiliki skala data yang sama. Berikut adalah contoh datanya

Tabel 4. 8 Contoh Data Hasil Normalisasi Fitur Interior

cut_Lengkap	cut_Sebagian	cut_Tak Berperabot
0	1	0
0	1	0
0	0	1
0	0	1
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	1	0
1	0	0
1	0	0
1	0	0
1	0	0
0	0	1
0	0	1
0	0	1
0	0	1
0	1	0
1	0	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

2. Menghitung Matrik Kovarian

Sebelum menghitung matrik kovarian, perlu diketahui matrik kovarian ini digunakan untuk menghitung variansi dan korelasi antar fitur dalam dataset. Nantinya nilai dari matrik varian kovarian berfungsi sebagai masukan untuk mendapatkankan nilai *eigen* dan *vektor eigen*.

Untuk menghitung matrik kovarian, terlebih dahulu untuk menghitung nilai varian dan nilai kovarian.

Varian (1 Atribut):

$$var(A_1) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})}{(n-1)} \quad (4.14)$$

Kovarian (2 Atribut):

$$cov(A_1, A_2) = \frac{\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})}{(n-1)} \quad (4.15)$$

Setelah diketahui varians dan kovariannya langkah selanjutnya yaitu menghitung matrik kovariannya. Berikut adalah rumusnya

$$C^{nxn} = (C_{i,j}), Where\ C_{i,j} = cov(A_1, A_2) \quad (4.16)$$

Berikut untuk nilai varian dari masing masing variabel

```
array([3.11944291e-01, 1.04570730e-01, 7.93555610e-32])
```

Gambar 4. 20 Nilai Varian Fitur Interior

Berikut adalah nilai dari matrik kovarian yang nantinya akan digunakan untuk menghitung *eigenvalue* dan *eigenvector*

```
array([[ 0.0748809 , -0.01473067, -0.06015023],
       [-0.01473067,  0.14810728, -0.13337661],
       [-0.06015023, -0.13337661,  0.19352684]])
```

Gambar 4. 21 Matrik Kovarian Fitur Interior

3. Menghitung *Eigenvector* dan *Eigenvalue*

Untuk menghitung *eigenvalue* dan eigenvektor diperlukan nilai dari kovarian matrik. Berikut adalah persamaan untuk menghitung *eigenvalue* dan *eigenvektor*.

$$Cv = \lambda v \qquad (4.\ 17)$$

Turunan dari rumus tersebut bisa menjadi seperti ini. M tersebut adalah nilai matrik kovarian yang telah dihitung ditahap sebelumnya.

$$(M - \lambda I)\, v = 0 \qquad (4.\ 18)$$

Setiap nilai *eigenvalue* atau "λ" harus memenuhi persamaan determinan. Setelah nilai *eigen value* ditemukan maka hasil dari *eigenvalue* dikembalikan dipersamaan $Cv = \lambda v$ untuk menghitung nilai *eigenvector*

$$|M - \lambda I| = 0 \qquad (4.\ 19)$$

Dari persamaan tersebut "v" adalah *eigenvector* dan "λ" adalah *eigenvalue*. Dari hasil perhitungan diperoleh nilai *eigenvalue* yang dihitung dari matriks kovariansi data yang telah diproses oleh PCA sebagai berikut

```
array([3.11944291e-01, 1.04570730e-01, 7.93555610e-32])
```

Gambar 4. 22 Matrik Kovarian Fitur interior

Untuk nilai *eigenvektor* sebagai berikut

```
array([[ 0.15785054,  0.61484156, -0.77269209],
       [ 0.80109293, -0.53724904, -0.26384389],
       [ 0.57735027,  0.57735027,  0.57735027]])
```

Gambar 4. 23 Nilai *Eigenvektor* Fitur Interior

4. Pemilihan *Eigenvector*

Setelah ketemu nilai *eigenvector* dan *eigenvalue*, langkah selanjutnya mengurutkan *eigenvector* dari *eigenvalue* yang paling besar ke yang paling kecil. Karena ada beberapa nilai *eigenvector* dari beberapa *eigenvalue*, maka perlu dihitung proporsi dari setiap *eigenvalue* mana yang nilai paling besar. *Eigenvalue* yang memiliki proporsi paling besar ini memiliki informasi yang paling tinggi, yang nantinya nilai *eigenvector* dari *eigenvalue* yang memiliki proporsi paling tinggi akan dijadikan pengali untuk mereduksi data.

Setelah menerapkan class PCA, langkah selanjutnya yaitu mengetahui proporsi informasi dari komponen.

```
array([0.749, 0.251, 0.   ])
```

Gambar 4. 24 Proporsi Informasi Komponen Fitur Interior

Arti dari output tersebut adalah 74,9% informasi terdapat pada PC pertama, 25,1% terdapat pada PC kedua, 0% pada PC ketiga.

5. Proyeksi Data

Pada hasil sebelumnya dapat dilihat pada PC pertama memiliki nilai tertinggi daripada yang lain. Maka dari itu maka akan mereduksi fitur dan hanya mempertahankan pada PC pertama saja. PC pertama ini akan menggantikan dari ketiga fitur 'cut_Lengkap','cut_Sebagian','cut_tak berperabot' menjadi fitur ‘Interior’. Berikut adalah hasil dari hasil reduksi

Tabel 4. 9 Contoh Hasil Reduksi Fitur Interior

Id	interior
1	1,060991
2	-0,32654
3	1,060991
4	-0,32654
5	-0,32654
6	-0,32654
7	-0,32654
8	-0,32654
9	1,060991
10	1,060991
11	-0,32654
12	-0,32654
13	1,060991
14	-0,32654
15	-0,32654
16	1,060991
17	-0,32654
18	-0,32654
19	-0,32654
20	-0,32654
21	-0,32654
22	1,060991
23	0,604
24	-0,32654
25	-0,32654

4.6.3 *Data Preprocessing – Train Test Split*

Tahap ini adalah tahap dimana data akan dibagi menjadi dua bagian. Tahap split data ini dilakukan untuk membagi jumlah dataset yang dimiliki menjadi data *training* dan data *testing*. Data *training* ini yang nantinya digunakan untuk melatih model dan data *testing* digunakan untuk melakukan tes terhadap model yang sudah dilatih.

Tahap split data ini dilakukan sebelum dilakukannya proses standarisasi pada data. Perlu mempertahankan sebagian data yang ada dalam kontek ini adalah data

testing untuk menguji seberapa generalisasi model dalam membaca data baru sebelum di normalisasi atau di standarisasi.

Dalam kasus penelitian ini, menggunakan perbandingan untuk melakukan split data yaitu 9 : 1. Perbandingan 9 untuk data *training* dan 1 untuk data *testing*. Data sebelum masuk ke tahap split data yang sudah melalui tahap *preprocessing* yaitu sebesar 4823 *rows*. Berikut adalah perbandingan untuk split data.

Tabel 4. 10 Split Data

Perbandingan	**Data *Training***	**Data *Testing***
9 : 1	2032	226

Data yang sudah melalui tahap split data tersebut akan masuk ketahap selanjutnya untk melakukan normalisasi data.

4.6.4 *Data Preprocessing* – Normalisasi

Machine learning akan memiliki performa yang lebih baik dan akan lebih cepat ketika dimodelkan dengan data yang mempunyai skala yang sama atau mendekati. Proses normalisasi ini akan membuat pengolahan yang dilakukan oleh algoritma *machine learning* ini akan lebih mudah. Normalisasi digunakan untuk menskalakan nilai agar cocok dalam rentang tertentu. Menyesuaikan rentang nilai sangat penting saat berhadapan dengan Atribut unit dan skala yang berbeda. Tahap normalisasi ini adalah tahap dimana nilai yang terdapat di data akan dirubah menjadi rentang nilai 0-1. Berikut adalah contoh data hasil normalisasi.

harga	kamar tidur	kamar mandi	luas tanah(...	harga per m	luas Bangun...	listrik	alamat	sertifikat
0.536	0.625	1	0.649	0.430	0.947	0.574	0	0
0.134	0.125	0.200	0.181	0.360	0.124	0.279	0.287	0
0.268	0.125	0.200	0.193	0.727	0.124	0.279	1	0
0.089	0.250	0	0.147	0.282	0.124	0.279	1	0
0.071	0.125	0	0.147	0.215	0.082	0.279	1	0
0.089	0.250	0	0.147	0.282	0.124	0.279	1	0
0.081	0.125	0	0.193	0.184	0.093	0.279	1	0
0.223	0.500	0.400	0.274	0.412	0.420	0.574	0.287	0
0.107	0.125	0	0.209	0.237	0.124	0.279	0.287	0
0.143	0.250	0.200	0.186	0.378	0.146	0.279	0.287	0
0.263	0.125	0.200	0.227	0.605	0.156	0.539	1	0
0.509	0.500	0.400	0.578	0.460	0.631	0.279	1	0
0.500	0.750	0.800	0.397	0.675	0.578	1	0.281	0

Gambar 4. 25 Contoh Data Normalisasi

4.7 Modeling

Algoritma *random forest* adalah salah satu algortima yang termasuk dalam *supervised learning*. Algoritma ini bisa menyelesaikan masalah klasifikasi ataupun regresi. Algoritma *Random Forest* termasuk dalam *ensemble learning*. *Ensemble learning* ini bekerja dengan beberapa metode yang bekerja sama untuk melakukan kinerjanya.

Terdapat dua hasil yang akan dijelaskan dipenelitian ini berdasarkan desain sistem yang telah dibuat diawal, yaitu hasil dari hasil prediksi yang hanya menggunakan *random forest* dan hasil yang menggunakan gabungan antara PCA dan *random forest*. Perbandingan hasil antara gabungan PCA dan *random forest* dengan model tanpa PCA ini dilakukan untuk melihat apakah penggunaan PCA dan *random forest* khususnya untuk kasus prediksi harga rumah dapat bekerja secara optimal atau tidak. Semua tahapan sebelum memasuki tahap modeling sama akan tetapi yang membedakan adalah penerapan PCA. PCA diterapkan dalam tahap *preprocessing* data, yaitu mereduksi fitur alamat dan sertifikat yang terpecah menjadi beberapa variabel baru.

4.8 Evaluasi

Untuk memastikan metode yang digunakan baik atau tidak yaitu dengan mengujinya. Untuk mengevaluasi model regresi secara teknis hanya menghitung selisih antara nilai sebenarnya dan nilai prediksi dalam hal ini bisa disebut error. Untuk menguji, matrik yang akan digunakan yaitu matrik RMSE, berikut adalah persamaanya

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(At - Ft)^2}{n}} \qquad (4.\ 20)$$

Dimana:

At = Nilai aktual/ nilai sebenarnya

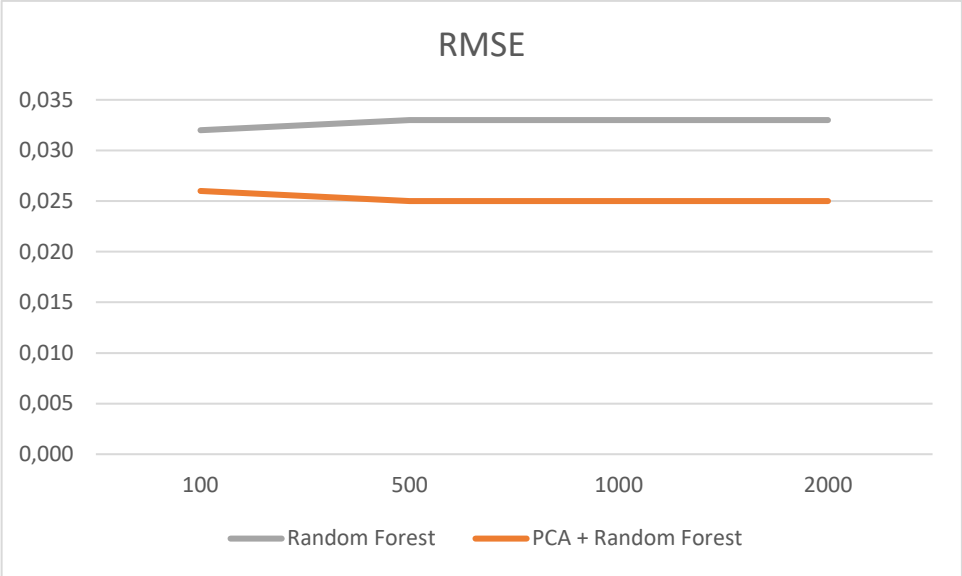
Ft = Nilai prediksi

N = Jumlah dataset

Kemudian dilakukan pengujian model dengan beberapa iterasi, dalam penelitian ini digunakan 4 iterasi yaitu 100, 500, 1000, dan 2000. Berikut adalah hasil dari hasil pengujian evaluasi model PCA dan *random forest*

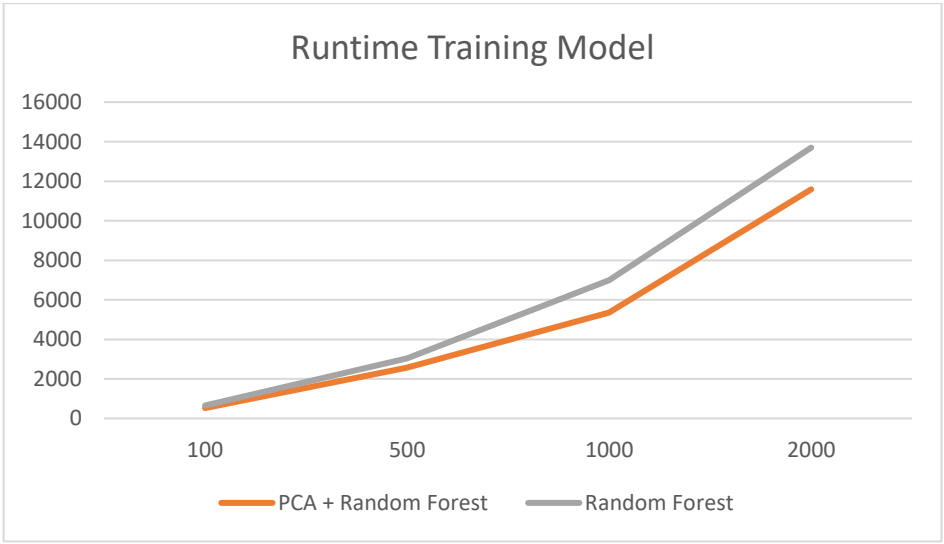
Tabel 4. 11 Hasil Pengujian

Iterasi	PCA + Random Forest		Random Forest	
	Error	Time	Error	Time
100	0,026	528	0,032	656
500	0,025	2563	0,033	3046
1000	0,025	5349	0,033	6997
2000	0,025	11587	0,033	13698
Average	0,0253	5007	0,03275	6099



Gambar 4. 26 RMSE

Dari gambar 4.21 dapat diketahui hasil evaluasi menggunakan nilai error yaitu menggunakan RMSE. Dari hasil visualisasi tersebut terdapat 4 iterasi yaitu 100, 500, 1000, 2000. Nilai error pada grafik *random forest* pada iterasi 100 memiliki nilai error 0.032, iterasi 500, 1000, 2000 memiliki nilai error 0.033. Pada grafik PCA dan random forest pada iterasi 100 memiliki nilai error 0.026, iterasi 500, 1000, 2000 memiliki nilai error 0.025.



Gambar 4. 27 *Runtime Training Model*

Dari gambar 4.22 dapat diketahui hasil waktu pelatihan model. Dari hasil visualisasi tersebut terdapat 4 iterasi yaitu 100, 500, 1000, 2000. Waktu pelatihan model pada grafik *random forest* pada iterasi 100 membutuhkan waktu 656 *milliseconds*, iterasi 500 membutuhkan waktu 3046 *milliseconds*, iterasi 1000 membutuhkan waktu 6997 *milliseconds*, iterasi 2000 membutuhkan waktu 13698 milliseconds. Pada grafik PCA dan *random forest* pada iterasi 100 membutuhkan waktu 528 *milliseconds*, iterasi 500 membutuhkan waktu 2563 *milliseconds*, iterasi 1000 membutuhkan waktu 5349 *milliseconds* dan iterasi 2000 membutuhkan waktu 11587 *milliseconds*.

Dari hasil pelatihan yang dapat dilihat di tabel 4.11, dapat disimpulkan bahwa penggunaan hasil evaluasi model yang menggunakan PCA memiliki tingkat error yang lebih kecil dan nilainya lebih konsisten yaitu dengan rata-rata 0.0253. Sedangkan hasil evaluasi tanpa PCA dan hanya menggunakan *Random Forest* memiliki nilai eror yang lebih besar yaitu dengan rata rata 0.03275. Waktu pelatihan menggunakan model PCA memiliki waktu yang lebih cepat yaitu dengan rata-rata 5007 *milisecond*, sedangkan yang hanya menggunakan *random forest* tanpa PCA memiliki waktu rata-rata sebesar 6099.

4.9 Integrasi Islam

4.9.1 HabluminAlloh

Rumah menjadi tempat dimana banyak rahmat Alloh diturunkan kepada setiap muslim. Dalam Al-Qur'an surah Al-Araf ayat 74 yang berbunyi

وَاذْكُرُوا إِذْ جَعَلَكُمْ خُلَفَاءَ مِنْ بَعْدِ عَادٍ وَبَوَّأَكُمْ فِي الْأَرْضِ تَتَّخِذُونَ مِنْ سُهُولِهَا قُصُورًا وَتَنْحِتُونَ الْجِبَالَ بُيُوتًا ۖ فَاذْكُرُوا آلَاءَ اللَّهِ وَلَا تَعْثَوْا فِي الْأَرْضِ مُفْسِدِينَ

Artinya: *"Dan ingatlah olehmu di waktu Tuhan menjadikam kamu pengganti-pengganti (yang berkuasa) sesudah kaum 'Aad dan memberikan tempat bagimu di bumi. Kamu dirikan istana-istana di tanah-tanahnya yang datar dan kamu pahat gunung-gunungnya untuk dijadikan rumah; maka ingatlah nikmat-nikmat Allah dan janganlah kamu merajalela di muka bumi membuat kerusakan." (QS. Al-Araf:74)*

Dalam Tafsir Jalalain dijelaskan bahwa Allah memberikan tempat bagi setiap muslim di bumi untuk dirikan istana (Rumah) dan Allah menyuruh untuk mengingat nikmat-nikmat Allah dan Allah menyuruh kita untuk tidak merusak apa yang disekitar kita.

Dijelaskan dalam tafsir dari kementrian agama bahwasannya orang-orang Samud juga diingatkan akan nikmat Allah untuk taat dan taat kepada-Nya. Dan ingatlah nikmat dan kebaikan Allah kepadamu ketika setelah kehancuran kaum 'ad, dia mengangkatmu menjadi khalifah yang kuat dan menempatkanmu di tempat yang memudahkan aktivitasmu di bumi, yaitu di tanah Hijriah, sebuah tempat yang strategis. Sabuk tempat tinggal Di musim panas mereka membangun istana, bangunan besar, luas dan indah di atas permukaan datar, yaitu tanah rendah. Dan di dataran tinggi, di perbukitan dan di bebatuan, Anda menggalinya dan melubanginya, sampai menjadi rumah untuk Anda tinggali selama musim dingin. Maka ingatlah nikmat Allah yang dilimpahkan kepadamu untuk bersyukur dan tidak berbuat

munkar di muka bumi dengan menyekutukan Allah, berbuat maksiat dan melalaikan dakwah utusan-Nya. Ketika mendengar peringatan Allah melalui Nabi yang saleh, para tokoh masyarakat yang angkuh dan angkuh serta murah hati umatnya berkata dengan nada mengejek untuk menimbulkan kecurigaan orang-orang yang dianggap lemah, atas nama orang-orang yang beriman di antara umatnya. Ketika kita percaya bahwa orang saleh adalah utusan Tuhannya yang diutus untuk menyampaikan perjanjiannya, orang-orang yang beriman menanggapi dengan penuh semangat. Padahal kami sangat mengimani apa yang beliau turunkan kepada kami yaitu perjanjian yaitu nabi yang saleh, karena petunjuk itu benar dan berasal dari Allah.

4.9.2 Habluminannas

Tujuan penelitian ini nantinya digunakan untuk membantu para pengembang dibidang properti untuk membantu proses bisnisnya. Perlu diketahui pihak pengembang juga harus berhati hati dalam menentukan harga properti. Perlu diketahui setiap tahunnya baik jangka pendek ataupun jangka panjang harga properti semakin naik dan bahkan hampir tidak pernah turun (Feng & Jones, 2015). Salah memperhitungkan harga properti membuat bisnis dan investasi bisa menjadi rugi. Resiko ini sangatlah tidak bagus jika dialami oleh setiap pebisnis. Tentu saja setiap bisnis selalu mengejar yang namanya profit. Penting bagi seorang pebisnis, terutama dibidang properti untuk mengetahui dan mampu untuk memprediksi harga properti. Sehingga pebisnis dapat mendapatkan profit sebesar mungkin.

Dalam hal ini proses membantu sesama muslim, khususnya para pengembang properti ini sesuai dengan hadist nabi terdapat dalam kitab Arba’in An Nawawi hadis ke 36 yang berbunyi

عَنْ أَبِي هُرَيرَة رَضِيَ اللهُ عَنْهُ عَنِ النَّبِيِّ ﷺ قَالَ: مَنْ نَفَّسَ عَنْ مُؤمِن كُربَةً مِن كُرَبِ الدُّنيَا نَفَّسَ اللهُ عَنهُ كُربَةً مِنْ كرَبِ يوم القيامَةِ، وَمَنْ يَسَّرَ على مُعْسِرٍ يَسَّرَ الله عَلَيهِ في الدُّنيَا والآخِرَة، وَمَنْ سَتَرَ مُسلِمَاً سَتَرَهُ الله في الدُّنيَا وَالآخِرَة، وَاللهُ في عَونِ العَبدِ مَا كَانَ العَبدُ في عَونِ أخيهِ، وَمَنْ سَلَكَ طَرِيقَاً يَلتَمِسُ فيهِ عِلمَاً سَهَّلَ اللهُ لهُ بِهِ طَرِيقَاً إِلَى الجَنَّةِ، وَمَا اجتَمَعَ قَومٌ في بَيتٍ مِنْ بيوتِ اللهِ يَتلونَ كِتابِ اللهِ وَيتَدارَسونَهَ بَينَهُم إِلا نَزَلَت عَلَيهُم السَّكينَة وَغَشيَتهم الرَّحمَة وحَفَتهُمُ المِلائِكة وَذَكَرهُم اللهُ فيمَن عِندَهُ،وَمَنْ بَطَّأ بِهِ عَمَلُهُ لَمْ يُسْرِعْ بِهِ نَسَبُهُ (رواه مسلم بهذا اللفظ)

Artinya: *"Abu Hurairah berkata bahwa Rasulullah bersabda, "Barang siapa yang menghilangkan sebuah kesulitan duniawi seorang mukmin, niscaya Allah akan menghilangkan darinya sebuah kesulitan pada hari kiamat. Barang siapa yang meringankan orang yang kesusahan, niscaya Allah akan meringankan baginya (kesusahannya) di dunia dan akhirat. Barang siapa yang menutupi aib seorang muslim, niscaya Allah akan menutupi aibnya di dunia dan akhirat. Allah akan senantiasa menolong hamba-Nya, selama hamba tersebut menolong saudaranya. Barang siapa yang menempuh satu jalan untuk mencari ilmu, niscaya Allah akan memudahkan baginya jalan menuju surga. Tidaklah suatu kaum berkumpul di salah satu rumah dari rumah–rumah Allah (masjid), membaca kitabullah, saling belajar di antara mereka, melainkan akan turun kepada mereka ketenangan, rahmat meliputi mereka, para malaikat mengerumuni mereka, serta Allah akan menyebut–nyebut mereka di hadapan makhluk yang berada di sisi-Nya. Barang siapa yang lambat dalam beramal, garis nasabnya tidak akan bisa membantunya."* (HR. Muslim dengan lafaz ini)

4.9.3 Habluminal'alam

Dalam surat Al-Araf ayat 74 yang berbunnyi

وَاذْكُرُوا إِذْ جَعَلَكُمْ خُلَفَاءَ مِنْ بَعْدِ عَادٍ وَبَوَّأَكُمْ فِي الْأَرْضِ تَتَّخِذُونَ مِنْ سُهُولِهَا قُصُورًا وَتَنْحِتُونَ الْجِبَالَ بُيُوتًا ۖ فَاذْكُرُوا آلَاءَ اللَّهِ وَلَا تَعْثَوْا فِي الْأَرْضِ مُفْسِدِينَ

Artinya: *"Dan ingatlah olehmu di waktu Tuhan menjadikam kamu pengganti-pengganti (yang berkuasa) sesudah kaum 'Aad dan memberikan tempat bagimu di bumi. Kamu dirikan istana-istana di tanah-tanahnya yang datar dan kamu pahat gunung-gunungnya untuk dijadikan rumah; maka ingatlah nikmat-nikmat Allah dan janganlah kamu merajalela di muka bumi membuat kerusakan." (QS. Al-Araf:74)*

Dijelaskan bahwasannya Alloh telah memberikan tempat bagi setiap muslim untuk mendirikan rumah. Dan dalam lanjutan ayat tersebut diikuti oleh larangan Alloh untuk tidak merusak alam apa yang ada disekitar kita.

Dalam ayat lain juga dijelaskan di surat Al-Hijr ayat 19 yang berbunyi sebagai berikut

وَٱلْأَرْضَ مَدَدْنَٰهَا وَأَلْقَيْنَا فِيهَا رَوَٰسِيَ وَأَنۢبَتْنَا فِيهَا مِن كُلِّ شَىْءٍ مَّوْزُونٍ

Artinya: *"Dan Kami telah menghamparkan bumi dan menjadikan padanya gunung-gunung dan Kami tumbuhkan padanya segala sesuatu menurut ukuran." (QS. Al-Hijr:19)*

Dari surat tersebut dapat dijelaskan bahwa setelah Alloh menjelaskan tanda kekuasaannya yang ada di langit dan bumi, Alloh menciptakan sesuatu menurut pada ukurannya sesuai dengan kemaslahatan makhluk-Nya. Alloh telah membagi sesuai dengan takarannya, dan Alloh melarang untuk merusak apa yang ada disekitar kita, sehingga apa yang ada disekitar kita tidak dirugikan dengan ketamaan, kerakusan.

Dalam tafsir Kementerian Agama, bahwa Usai menyebutkan tanda kekuasaannya di langit, Tuhan kemudian menyebutkan tanda kekuasaannya di bumi. Allah berfirman dan Kami bentangkan bumi sebagai alas manusia dan Kami bangun di atasnya gunung-gunung yang kuat untuk bumi agar tidak runtuh dan goncang sampai manusia aman dan Kami menciptakan dan menumbuhkan apa yang ada di dalamnya, seperti berbagai tanaman, menurut sebuah takaran seimbang dan tepat; semua untuk kebaikan makhluknya. Dan selain itu Kami juga telah menciptakan sarana dan gaya hidup untuk kebutuhanmu wahai manusia, baik berupa sandang, pangan dan piring. Dan kami juga telah menciptakan berbagai makhluk yang bukan kamu pemberi makanannya, tetapi kami yang membawanya.

BAB V
PENUTUP

5.1 Kesimpulan

Hasil analis dapat disimpulkan bahwa penjualan rumah terbanyak terdapat di daerah kecamatan lowokwaru, dan rumah yang mempunya sertifikat SHM – Sertifikat Hak Milik memiliki penjualan yang tinggi. Dari beberapa variabel bahwa variabel atau fitur luas tanah dan luas bangunan memiliki pengaruh yang tinggi terhadap harga, sedangkan variabel listrik memiliki korelasi yang rendah. Kemudian untuk hasil pelatihan model dapat disimpulkan bahwa:

1. Penggunaan hasil evaluasi model yang menggunakan PCA memiliki tingkat error yang lebih kecil dan nilainya lebih konsisten yaitu dengan rata-rata 0.0253. Sedangkan hasil evaluasi tanpa PCA dan hanya menggunakan Random Forest memiliki nilai eror yang lebih besar yaitu dengan rata rata 0.03275. Waktu pelatihan menggunakan model PCA memiliki waktu yang lebih cepat yaitu dengan rata-rata 5007 milisecond, sedangkan yang hanya menggunakan random forest tanpa PCA memiliki waktu rata-rata sebesar 6099.
2. Penggunaan PCA dan *Random Forest* memiliki hasil yang lebih optimal dibandingkan dengan yang hanya menggunakan *Random Forest*. Hal ini bisa dilihat dari perbandingan hasil evaluasi model.

5.2 Saran

Peneliti mengakui bahwa penelitian ini masih perlu untuk dikembangkan, adapun saran untuk penelitian selanjutnya sebagai berikut

1. Perlu menggunakan dataset dari sumber lain atau daerah lain yang ada di indonesia dengan lebih banyak variabelnya dan jumlah datasetnya.
2. Ada banyak metode yang bisa digunakan untuk melakukan prediksi harga rumah ini, dari literature review diperoleh banyak metode yang memiliki akurasi yang tinggi atau tingkat error yang rendah yang bisa digunakan.

DAFTAR PUSTAKA

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Breiman, L. (2001). Random Forests—Machine Learning. *Kluwer Academic Publishers*, *45*(1), 5–32.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, *7*(5), 168. https://doi.org/10.3390/ijgi7050168

De Nadai, M., & Lepri, B. (2018). The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 323–330. https://doi.org/10.1109/DSAA.2018.00043

Durganjali, P., & Pujitha, M. V. (2019). House Resale Price Prediction Using Classification Algorithms. *2019 International Conference on Smart Structures and Systems (ICSSS)*, 1–4. https://doi.org/10.1109/ICSSS.2019.8882842

El Boujnouni, H., Rahouti, M., & El Boujnouni, M. (2021). Identification of SARS-CoV-2 origin: Using Ngrams, principal component analysis and Random Forest algorithm. *Informatics in Medicine Unlocked*, *24*, 100577. https://doi.org/10.1016/j.imu.2021.100577

Feng, Y., & Jones, K. (2015). Comparing multilevel modelling and artificial neural networks in house price prediction. *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, 108–114. https://doi.org/10.1109/ICSDM.2015.7298035

Fuentes, A. (2018). *Hands-On Predictive Analytics with Python*. Packt Publishing.

Gardner, C., & Lo, D. C.-T. (2021). PCA Embedded Random Forest. *SoutheastCon 2021*, 1–6. https://doi.org/10.1109/SoutheastCon45413.2021.9401949

Garriga, C., Hedlund, A., Tang, Y., & Wang, P. (2021). Rural-urban migration and house prices in China. *Regional Science and Urban Economics*, *91*, 103613. https://doi.org/10.1016/j.regsciurbeco.2020.103613

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2004). Random forest classification of multisource remote sensing and geographic data. *IEEE*

International IEEE International IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004, *2*, 1049–1052. https://doi.org/10.1109/IGARSS.2004.1368591

Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW. *PLANNING MALAYSIA*, *19*. https://doi.org/10.21837/pm.v19i17.1018

Jiang, P., Sun, X., & Lu, Z. (2007). Quantitative Estimation of siRNAs Gene Silencing Capability by Random Forest Regression Model. *2007 1st International Conference on Bioinformatics and Biomedical Engineering*, 230–233. https://doi.org/10.1109/ICBBE.2007.62

Jiang, Z., & Shen, G. (2019). Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework. *2019 6th International Conference on Systems and Informatics (ICSAI)*, 1408–1412. https://doi.org/10.1109/ICSAI48974.2019.9010071

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, *111*, 104919. https://doi.org/10.1016/j.landusepol.2020.104919

Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural networks. *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 518–522. https://doi.org/10.1109/FSKD.2016.7603227

Lu, S., Li, Q., Yu, H., & Wang, X. (2020). Damage Evaluation Method of CFRP Structures Based on PCA and Random Forest Algorithm. *2020 Chinese Automation Congress (CAC)*, 3804–3807. https://doi.org/10.1109/CAC51589.2020.9327009

Madhuri, CH. R., Anuradha, G., & Pujitha, M. V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. *2019 International Conference on Smart Structures and Systems (ICSSS)*, 1–5. https://doi.org/10.1109/ICSSS.2019.8882834

Nguyen, T.-T., Huang, J. Z., & Nguyen, T. T. (2015). Unbiased Feature Selection in Learning Random Forests for High-Dimensional Data. *The Scientific World Journal*, *2015*, 1–18. https://doi.org/10.1155/2015/471371

Nur, A., Ema, R., Taufiq, H., & Firdaus, W. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications*, *8*(10). https://doi.org/10.14569/IJACSA.2017.081042

Phan, T. D. (2018). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (ICMLDE)*, 35–42. https://doi.org/10.1109/iCMLDE.2018.00017

R.M.A. van der Schaar. (2015). Analisis Pasar Properti Indonesia; Overview & Kepemilikan Asing. *Indonesia Investment*. Analysis of Indonesian Property Market; Overview and Foreign Ownership

Shahhosseini, M., Hu, G., & Pham, H. (2020a). Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction. In H. Yang, R. Qiu, & W. Chen (Eds.), *Smart Service Systems, Operations Management, and Analytics* (pp. 87–97). Springer International Publishing. https://doi.org/10.1007/978-3-030-30967-1_9

Shahhosseini, M., Hu, G., & Pham, H. (2020b). Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction. In H. Yang, R. Qiu, & W. Chen (Eds.), *Smart Service Systems, Operations Management, and Analytics* (pp. 87–97). Springer International Publishing. https://doi.org/10.1007/978-3-030-30967-1_9

Sharma, N., Arora, Y., Makkar, P., Sharma, V., & Gupta, H. (2021). Real Estate Price's Forecasting Through Predictive Modelling. In A. Joshi, M. Khosravy, & N. Gupta (Eds.), *Machine Learning for Predictive Analysis* (Vol. 141, pp. 589–597). Springer Singapore. https://doi.org/10.1007/978-981-15-7106-0_58

Smallman, L., Artemiou, A., & Morgan, J. (2018). Sparse Generalised Principal Component Analysis. *Pattern Recognition*, *83*, 443–455. https://doi.org/10.1016/j.patcog.2018.06.014

Song, Q., & Huang, Y. (2021). A Solution for Liquor Recognition Based on PCA-RF and Laser Induced Fluorescence. *IEEE Access*, *9*, 35101–35108. https://doi.org/10.1109/ACCESS.2021.3049941

Waskle, S., Parashar, L., & Singh, U. (2020). Intrusion Detection System Using PCA with Random Forest Approach. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 803–808. https://doi.org/10.1109/ICESC48915.2020.9155656

Wiradinata, T., Graciella, F., Tanamal, R., Soekamto, Y. S., & Saputri, T. R. D. (2022). POST-PANDEMIC ANALYSIS OF HOUSE PRICE PREDICTION IN SURABAYA: A MACHINE LEARNING APPROACH. *Journal of Southwest Jiaotong University*, *57*(5), 562–573. https://doi.org/10.35741/issn.0258-2724.57.5.45

LAMPIRAN

Contoh Sample Data Sebelum Tahap PCA

kamar tidur	kamar mandi	luas tanah(m2)	harga per m	luas Bangunan(m2)	parkir	listrik	harga	cut_Blimbing	cut_Kedungkandan	cut_Klojen	cut_Lowokwaru	cut_Sukun	cut_AJB	cut_SHGB	cut_SHM	cut_Lengkap	cut_Sebagian	cut_Tak
6	6	291	10,30928	450	1	2200	3000000000	1	0	0	0	0	0	0	1	0	1	0
2	2	90	16,66667	60	1	1300	1500000000	0	0	0	1	0	0	0	1	0	1	0
2	1	70	5,714286	40	1	1300	400000000	0	0	0	1	0	0	0	1	0	0	1
3	1	70	7,142857	60	1	1300	500000000	0	0	0	1	0	0	0	1	0	0	1
3	3	126	19,84127	100	1	2200	2500000000	0	1	0	0	0	0	0	1	0	1	0
3	2	325	6,461538	300	1	1300	2100000000	1	0	0	0	0	0	0	1	0	0	1
2	1	83	7,46988	70	1	1300	620000000	0	1	0	0	0	0	0	1	0	0	1
3	1	75	5,933333	49	1	1300	445000000	0	1	0	0	0	0	0	1	0	0	1
2	1	72	8,041667	40	1	1300	579000000	0	0	0	1	0	1	0	0	0	0	1
3	1	127	5,511811	60	1	1300	700000000	1	0	0	0	0	0	0	1	0	1	0
8	4	84	11,30952	100	1	1300	950000000	0	1	0	0	0	0	0	1	1	0	0
3	1	165	4,969697	100	1	1300	820000000	0	0	0	1	0	0	0	1	1	0	0
4	2	200	6,25	180	2	2200	1250000000	0	0	0	1	0	0	0	1	1	0	0

2	1	72	12,15278	43	1	1300	875000000	0	0	0	0	1	0	0	1	1	0	0
2	1	228	12,7193	100	1	2200	2900000000	0	0	1	0	0	0	0	1	0	0	1
2	1	228	12,7193	100	1	2200	2900000000	0	0	1	0	0	0	0	1	0	0	1
1	1	66	5,757576	66	1	1300	380000000	1	0	0	0	0	1	0	0	0	0	1
2	1	28	3,571429	20	1	900	100000000	0	0	0	1	0	1	0	0	0	0	1
3	1	102	5,392157	102	1	450	550000000	1	0	0	0	0	1	0	0	0	1	0
3	2	95	15,26316	75	1	1300	1450000000	0	0	0	1	0	1	0	0	1	0	0
2	1	75	4,666667	75	1	1300	350000000	1	0	0	0	0	1	0	0	0	0	1
2	1	60	4,333333	42	1	900	260000000	1	0	0	0	0	1	0	0	0	0	1
2	1	72	4,861111	50	1	1300	350000000	1	0	0	0	0	1	0	0	0	0	1
3	2	105	14,28571	120	1	1300	1500000000	0	0	0	1	0	0	1	0	0	0	1
2	1	78	8,653846	36	1	1300	675000000	0	0	0	0	1	0	1	0	0	0	1
2	1	78	8,653846	36	1	1300	675000000	0	0	0	0	1	0	1	0	0	0	1
2	1	105	6,190476	70	1	1300	650000000	0	0	0	1	0	0	1	0	0	0	1

Contoh Sample Data Setelah Tahap PCA

kamar tidur	kamar mandi	luas tanah(m2)	harga per m	luas Bangunan(m2)	parkir	listrik	harga	alamat	sertifikat	interior
6	6	291	10,30928	450	1	2200	3000000000	-0,6405	-0,10403	1,060991
2	2	85	8,823529	60	1	1300	750000000	-0,24715	-0,10403	-0,32654
2	2	90	16,66667	60	1	1300	1500000000	0,701776	-0,10403	1,060991
2	1	70	5,714286	40	1	1300	400000000	0,701776	-0,10403	-0,32654
3	1	70	7,142857	60	1	1300	500000000	0,701776	-0,10403	-0,32654
2	1	70	5,714286	40	1	1300	400000000	0,701776	-0,10403	-0,32654
3	1	70	7,142857	60	1	1300	500000000	0,701776	-0,10403	-0,32654
2	1	90	5,055556	45	1	1300	455000000	0,701776	-0,10403	-0,32654
5	3	126	9,920635	200	1	2200	1250000000	-0,24715	-0,10403	1,060991
2	1	97	6,185567	60	1	1300	600000000	-0,24715	-0,10403	1,060991
3	2	87	9,195402	70	1	1300	800000000	-0,24715	-0,10403	-0,32654
5	3	260	10,96154	300	2	1300	2850000000	0,701776	-0,10403	-0,32654
3	3	126	19,84127	100	1	2200	2500000000	-0,31396	-0,10403	1,060991
3	2	325	6,461538	300	1	1300	2100000000	-0,6405	-0,10403	-0,32654
3	2	140	5,714286	100	1	1300	800000000	0,701776	-0,10403	-0,32654
3	2	102	14,21569	120	1	2200	1450000000	0,701776	-0,10403	1,060991
2	1	72	4,513889	36	1	1300	325000000	-0,31396	-0,10403	-0,32654
2	1	60	6,65	45	1	1300	399000000	-0,31396	-0,10403	-0,32654
2	1	83	7,46988	70	1	1300	620000000	-0,31396	-0,10403	-0,32654
3	1	75	5,933333	49	1	1300	445000000	-0,31396	-0,10403	-0,32654
2	1	72	8,041667	40	1	1300	579000000	0,701776	0,62773	-0,32654