

LAPORAN PROYEK MATA KULIAH

12S4054 – DATA MINING

House Prices
Advanced Regression Techniques using XGboost



Disusun Oleh:

12S21009	Mikhael Janugrah Pakpahan
12S21010	Bobby Willy Siagian
12S21011	Aldi Jeremy Simamora

Tautan GitHub : <https://github.com/MikhaelJP/project-DaMi-03>

PROGRAM STUDI SARJANA SISTEM INFORMASI
FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO
INSTITUT TEKNOLOGI DEL
2024

DAFTAR ISI

BAB I PENDAHULUAN	5
1.1 Latar Belakang.....	5
1.2 Tujuan.....	6
1.3 Manfaat.....	6
1.4 Ruang Lingkup	6
1.5 Istilah dan Singkatan.....	7
BAB II STUDI LITERATURE.....	8
2.1 <i>Regresi</i>	8
2.1.1 Linear Regression.....	8
2.1.2 Ridge and Lasso Regression.....	8
2.1.3 XGboost	9
2.2 Prediksi Harga Rumah.....	
2.3 CRISP-DM	
BAB III METODE.....	10
3.1 CRISP-DM.....	10
3.1.1 Business Understanding	10
3.1.2 Data Understanding.....	11
3.1.3 Data Preparation	23
3.1.4 Modelling	33
3.1.5 Evaluation.....	37
3.1.6 Deployment	41
3.2 Timeline.....	46
BAB IV KESIMPULAN DAN SARAN.....	47
4.1 Kesimpulan.....	47
BAB V PEMBAGIAN PEKERJAAN	48
REFERENSI.....	49
LAMPIRAN	51

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam dunia properti, penentuan harga rumah yang akurat memiliki dampak yang signifikan terhadap berbagai pemangku kepentingan, termasuk pembeli, penjual, investor, dan pengembang properti. Penentuan harga yang tidak akurat dapat menyebabkan ketidakpuasan pelanggan, penurunan kepercayaan pasar, dan risiko finansial bagi pelaku usaha di sektor properti. Oleh karena itu, pengembangan model prediksi harga rumah yang efektif menjadi sangat penting untuk mendukung pengambilan keputusan yang lebih baik.

Teknik regresi lanjutan seperti **XGBoost (Extreme Gradient Boosting)** telah menjadi salah satu alat yang paling populer dalam analisis data dan pembelajaran mesin. XGBoost adalah algoritma boosting berbasis pohon keputusan yang terkenal karena kecepatan, efisiensi, dan akurasi prediksinya. Algoritma ini menggunakan teknik optimasi untuk mengurangi kesalahan prediksi, menjadikannya pilihan utama untuk banyak kompetisi pembelajaran mesin, seperti yang diselenggarakan oleh Kaggle ([Chen & Guestrin, 2016](#)). Pada proyek ini, kami akan memanfaatkan dataset **"House Prices: Advanced Regression Techniques"** dari Kaggle. Dataset ini berisi lebih dari 80 variabel yang mencakup berbagai fitur rumah, seperti ukuran, lokasi, kondisi, dan lainnya. Tantangan utama adalah memprediksi harga jual rumah berdasarkan fitur-fitur tersebut.

Selain itu, teknik XGBoost memungkinkan pemrosesan data dengan jumlah fitur yang besar dan memberikan kemampuan interpretabilitas melalui pentingnya fitur (feature importance). Hal ini menjadikannya alat yang ideal untuk memecahkan masalah prediksi harga rumah yang kompleks.

Proyek ini memiliki relevansi dalam konteks praktis dan akademik. Dalam konteks praktis, model prediksi harga rumah dapat digunakan oleh agen real estat, pembeli, dan pengembang untuk memahami pasar properti dengan lebih baik dan membuat keputusan yang lebih baik. Dalam konteks akademik, proyek ini menawarkan peluang untuk mengeksplorasi teknik pembelajaran mesin modern, seperti XGBoost, dalam penerapan dunia nyata, sehingga memperkaya literatur penelitian terkait metode regresi lanjutan.

1.2 Tujuan

Adapun tujuan dari pembuatan proyek ini adalah:

1. Salah satu syarat penyelesaian mata kuliah Data Mining T.A. 2024/2025
2. Mengembangkan model prediksi harga rumah yang akurat menggunakan XGBoost
3. Mengeksplorasi dan memahami hubungan antara fitur-fitur dalam dataset dengan harga rumah.

1.3 Manfaat

Adapun manfaat dari proyek ini adalah sebagai berikut:

1. Mengembangkan keterampilan dalam machine learning untuk memprediksi harga rumah secara akurat.
2. Membantu pemangku kepentingan, seperti agen real estat dan pengembang properti, dalam pengambilan keputusan berbasis data.
3. Menggunakan hasil prediksi sebagai dasar untuk memahami faktor-faktor yang memengaruhi harga rumah.

1.4 Ruang Lingkup

Ruang lingkup dari proyek yang kami kerjakan adalah:

1. Siklus pengerjaan mengikuti CRISP-DM yang terdiri dari tahap business understanding, data understanding, data preparation, modeling, evaluation, dan deployment.
2. Dataset yang digunakan untuk melatih model adalah dataset "House Prices: Advanced Regression Techniques" dari Kaggle yang terdiri dari lebih dari 80 fitur dan ribuan baris data.

1.5 Istilah dan Singkatan

Tabel 1. Istilah dan Singkatan

Singkatan	Definisi
AI	R2 Score
CRISP-DM	Proses yang digunakan untuk manajemen proyek data mining.
XGBoost	Algoritma machine learning yang merupakan implementasi dari teknik boosting berbasis pohon keputusan yang dirancang untuk meningkatkan performa model serta efisiensi komputasi.
MAE	Mean Absolute Error, metrik yang digunakan untuk mengevaluasi performa model regresi dengan menghitung rata-rata nilai absolut dari selisih prediksi dan nilai aktual.
RMSE	Root Mean Squared Error, metrik evaluasi yang mengukur kesalahan prediksi dengan memberi bobot lebih besar pada kesalahan yang lebih besar.
R2 Score	Koefisien determinasi yang mengukur seberapa baik model dapat menjelaskan variabilitas data target.

BAB II

STUDI LITERATURE

2.1 Regresi

Regresi adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel independen (fitur) dan variabel dependen (target). Dalam konteks pembelajaran mesin, regresi digunakan untuk memprediksi nilai kontinu, seperti harga rumah, berdasarkan atribut tertentu. Metode regresi mencakup berbagai pendekatan, mulai dari linear regression yang sederhana hingga metode non-linear dan kompleks seperti XGBoost. Regresi sangat berguna untuk memahami faktor-faktor yang memengaruhi variabel target serta membuat prediksi yang akurat ([Montgomery et al., 2012](#)).

2.1.1 Linear Regression

Linear regression adalah teknik dasar dalam regresi yang memodelkan hubungan linier antara variabel input dan output. Model ini sederhana, cepat, dan memberikan interpretabilitas yang baik, namun tidak efektif untuk menangani hubungan non-linier dalam dataset. Dalam kasus prediksi harga rumah, linear regression dapat digunakan untuk mendeteksi hubungan langsung antara fitur seperti luas tanah atau jumlah kamar dengan harga rumah ([Montgomery et al., 2012](#)).

2.1.2 Ridge dan Lasso Regression

Ridge dan Lasso Regression adalah metode regresi yang menggunakan regularisasi untuk mengurangi overfitting dan menangani multikolinearitas antar fitur. Ridge regression menambahkan penalti pada besar koefisien, sementara Lasso regression memungkinkan seleksi fitur secara otomatis dengan menekan koefisien fitur yang tidak relevan menjadi nol (Tibshirani, 1996).

2.1.3 XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning berbasis pohon keputusan yang mengimplementasikan teknik boosting untuk meningkatkan akurasi prediksi. Algoritma ini dikenal dengan efisiensi komputasi, kecepatan, dan kemampuan menangani dataset dengan jumlah fitur yang besar. XGBoost menggunakan pendekatan regularisasi L1 dan L2 untuk mengurangi overfitting dan menyediakan feature importance, sehingga cocok untuk tugas prediksi kompleks seperti harga rumah ([Chen & Guestrin, 2016](#)).

2.2 Prediksi Harga Rumah

Prediksi harga rumah adalah salah satu aplikasi populer dalam pembelajaran mesin, yang bertujuan untuk memperkirakan harga properti berdasarkan fitur tertentu. Beberapa faktor yang sering digunakan meliputi ukuran rumah, jumlah kamar, lokasi, dan usia properti. Analisis harga rumah penting untuk membantu agen real estat, pembeli, dan pengembang dalam pengambilan keputusan berbasis data. Dalam penelitian sebelumnya, algoritma seperti linear regression, random forest, dan XGBoost telah digunakan secara luas untuk memprediksi harga rumah dengan hasil yang signifikan ([Kaggle, 2024](#)).

2.3 CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah kerangka kerja yang digunakan untuk pengembangan proyek data mining dan pembelajaran mesin. Proses ini terdiri dari enam tahap utama, yaitu:

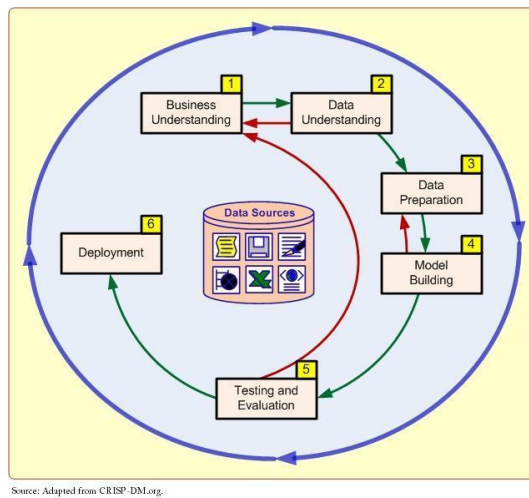
1. **Business Understanding** – Mendefinisikan tujuan proyek dan kebutuhan bisnis.
2. **Data Understanding** – Mengeksplorasi dataset untuk memahami struktur dan kualitas data.
3. **Data Preparation** – Membersihkan dan mempersiapkan data untuk modeling.
4. **Modeling** – Melatih model pembelajaran mesin, seperti XGBoost, untuk membuat prediksi.
5. **Evaluation** – Mengevaluasi kinerja model menggunakan metrik seperti RMSE, MAE, atau R2.
6. **Deployment** – Mengimplementasikan model dalam lingkungan nyata (Wirth & Hipp, 2000).

BAB III

METODE DAN ANALISIS

3.1 CRISP-DM

CRISP-DM adalah salah satu siklus pengerjaan proyek dalam Data Mining. Model ini mencakup tahap pengerjaan, apa yang dilakukan pada setiap tahap, dan hasil dari setiap tahap tersebut. Siklus ini dapat dilihat pada Gambar 1.



Gambar 1. CRISP-DM Step

Urutan yang ditunjukkan oleh tanda panah pada gambar tidak kaku, tanda panah hanya menunjukkan keterikatan antar tahapan. Metode ini akan kami gunakan dalam pengerjaan proyek Data Mining kami, setiap tahapan pada siklus akan dijelaskan pada sub-bab berikutnya.

3.1.1 Business Understanding

Tujuan utama proyek ini adalah untuk mengembangkan sebuah aplikasi dunia nyata, khususnya dalam konteks kesehatan mental. Beberapa masalah yang ingin diselesaikan adalah,

1. Depresi adalah masalah kesehatan mental yang signifikan di masyarakat, dan deteksi dini dapat menjadi langkah penting untuk memberikan perawatan yang tepat waktu.
2. Mengingat stigma yang masih ada terkait dengan kesehatan mental, alat deteksi yang efektif dan efisien dapat memberikan dampak positif dalam meningkatkan kesadaran individu agar lebih peka dengan kesehatan mental.
3. Dengan menggunakan algoritma XGBoost untuk memproses data, proyek ini berfokus pada penggunaan data besar (big data) dan kecerdasan buatan (AI) untuk memberikan solusi yang dapat diterapkan di dunia nyata, yaitu dalam bidang kesehatan mental.

3.1.2 Data Understanding

Tahap Data Understanding berfokus pada eksplorasi, pemeriksaan, dan pengumpulan wawasan awal dari data yang akan digunakan untuk menyelesaikan masalah yang telah didefinisikan. Tujuannya adalah untuk memahami karakteristik data secara menyeluruh dan memastikan bahwa data tersebut sesuai untuk analisis lebih lanjut.

3.1.2.1 DETERMINE BUSINESS OBJECTIVE

Dalam konteks analisis harga rumah, pemahaman bisnis (business understanding) menjadi langkah pertama yang sangat penting untuk membangun solusi yang tepat guna. Penentuan harga rumah dalam industri properti melibatkan berbagai faktor yang kompleks dan dinamis, yang tidak hanya berkaitan dengan nilai ekonomi dari sebuah properti, tetapi juga dipengaruhi oleh kondisi pasar, lokasi, karakteristik fisik rumah, serta faktor lingkungan dan sosial yang lebih luas. Bagi agen properti, pemilik rumah, investor, dan calon pembeli, harga rumah menjadi salah satu parameter utama yang digunakan untuk pengambilan keputusan. Oleh karena itu, adanya metode yang dapat memprediksi harga rumah dengan akurasi yang lebih tinggi sangat penting untuk meningkatkan efisiensi dan mendukung keputusan yang lebih cerdas.

Secara tradisional, penentuan harga rumah sering kali mengandalkan analisis manual yang dilakukan oleh agen atau evaluator properti. Proses ini biasanya memerlukan pengetahuan pasar lokal yang mendalam dan pemahaman terhadap berbagai faktor yang mempengaruhi harga, yang sering kali bersifat subjektif dan rentan terhadap kesalahan. Selain itu, fluktuasi pasar yang dipengaruhi oleh kondisi ekonomi, kebijakan pemerintah, dan tren pasar yang terus berubah, semakin menambah kompleksitas dalam proses ini.

Untuk mengatasi tantangan ini, pendekatan berbasis machine learning menawarkan solusi yang lebih otomatis dan akurat. Dalam hal ini, algoritma XGBoost (eXtreme Gradient Boosting) menjadi pilihan yang sangat tepat untuk digunakan dalam prediksi harga rumah. XGBoost adalah metode pembelajaran mesin yang menggunakan teknik boosting, di mana sejumlah model pohon keputusan yang relatif sederhana (weak learners) digabungkan untuk membentuk sebuah model yang kuat dan akurat. Dengan kemampuan untuk menangani masalah regresi dengan baik, XGBoost mampu memprediksi harga rumah dengan lebih presisi, memperbaiki kesalahan prediksi secara iteratif, dan memberikan hasil yang lebih stabil serta konsisten dibandingkan dengan metode konvensional.

Pemahaman bisnis dalam proyek ini adalah untuk memastikan bahwa model prediksi harga rumah tidak hanya memberikan akurasi tinggi, tetapi juga dapat memberikan wawasan yang dapat dipahami oleh pemangku kepentingan industri properti. Dengan demikian, proyek ini tidak hanya bertujuan untuk menciptakan solusi teknis yang efisien, tetapi juga untuk memastikan bahwa hasil prediksi yang diberikan relevan, transparan, dan dapat digunakan untuk mendukung keputusan strategis di pasar properti yang sangat dinamis.

3.1.2.2 DETERMINE PROJECT GOAL

Tujuan utama dari proyek ini adalah mengembangkan model prediksi menggunakan algoritma XGBoost (eXtreme Gradient Boosting), yang terbukti efektif untuk masalah regresi terutama pada data terstruktur seperti dataset harga rumah. Model ini akan dilatih dengan data historis tentang properti dan harga jualnya, dengan sasaran untuk menghasilkan prediksi harga yang akurat untuk properti baru yang belum pernah muncul dalam data sebelumnya. Teknik XGBoost dipilih karena

kemampuannya menangani fitur kompleks dan keandalannya dalam meningkatkan akurasi prediksi melalui metode boosting.

3.1.2.2 PRODUCE PROJECT PLAN

Berikut adalah jadwal pengerjaan proyek yang akan dilakukan:

Table 1. Project Plan

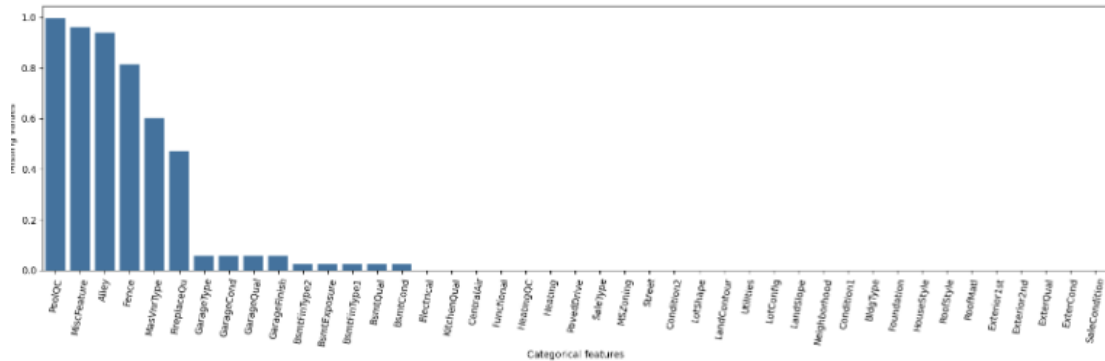
Tahapan	Waktu Pengerjaan	Kegiatan
<i>Business Understanding</i>	3 hari	Pada tahap ini akan dilakukan penentuan terhadap objektif bisnis, menentukan tujuan proyek, dan membuat rencana proyek.
<i>Data Understanding</i>	3 hari	Pada tahap ini data yang akan digunakan akan dikumpulkan, dianalisis, dan divalidasi.
<i>Data Preparation</i>	4 hari	Pada tahap ini akan dilakukan <i>data cleaning</i> , <i>data construction</i> , dan <i>data integration</i> .
<i>Modeling</i>	3 hari	Pada tahap ini akan dilakukan pemodelan terhadap model serta pengujian model.
<i>Evaluation</i>	3 hari	Pada tahap ini akan dilakukan evaluasi terhadap hasil model yang dibangun dan melakukan review terhadap proses pembangunan model.
<i>Deployment</i>	4 hari	Pada tahap ini akan dilakukan <i>deployment</i> , <i>monitoring</i> , dan <i>maintenance</i> terhadap model.

Dalam pelaksanaan proyek ini, Python menjadi salah satu alat penting yang akan digunakan. Python adalah bahasa pemrograman serbaguna yang mampu menjalankan berbagai perintah secara langsung melalui metode berorientasi objek dan menggunakan sintaks yang mudah dibaca. Python terkenal dalam bidang Data Science, Machine Learning, dan Internet of Things (IoT), menjadikannya pilihan utama untuk analisis data dan pembuatan model prediksi. Pada proyek ini, algoritma utama yang digunakan adalah XGBoost (Extreme Gradient Boosting). XGBoost merupakan algoritma machine learning yang menggunakan pendekatan ensemble learning, mirip dengan metode Random Forest. Namun, XGBoost memiliki fokus yang berbeda, yaitu menghasilkan model pohon keputusan secara bertahap untuk meningkatkan kinerja dan akurasi prediksi. Dalam XGBoost, setiap pohon (tree) dibangun secara berurutan, dengan bobot khusus yang diberikan pada masing-masing pohon untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. Dengan strategi ini, XGBoost mampu menangkap pola yang lebih kompleks dalam data, membuatnya sangat efektif untuk prediksi harga rumah.

3.1.2.3 Missing Values

Missing values (nilai yang hilang) adalah data yang tidak tercatat atau kosong dalam dataset, baik karena tidak tersedia, tidak dicatat dengan benar, atau terjadi kesalahan selama pengumpulan atau pengolahan data. Diagram pada Gambar 2. adalah visualisasi missing values pada dataset, di mana setiap kolom dataset direpresentasikan oleh batang vertikal.

Interpretasi Tinggi Batang



Gambar 2. Missing Values

Gambar 2 di atas menunjukkan sebuah bar chart yang menggambarkan importance score atau tingkat kepentingan dari fitur-fitur kategori dalam sebuah model machine learning. Pada sumbu X, terdapat berbagai fitur kategori seperti *PoolQC*, *MiscFeature*, *Alley*, *Fence*, dan lainnya, sementara sumbu Y merepresentasikan nilai kepentingan masing-masing fitur. Dari visualisasi ini, terlihat bahwa fitur *PoolQC*, *MiscFeature*, dan *Alley* memiliki importance score tertinggi, mendekati nilai 1, yang menunjukkan bahwa fitur-fitur tersebut memiliki pengaruh yang sangat signifikan terhadap performa model.

Di sisi lain, fitur seperti *Fence*, *MasVnrType*, dan *FireplaceQu* juga memiliki kontribusi yang cukup signifikan, meskipun lebih rendah dibandingkan tiga fitur teratas. Sebaliknya, sebagian besar fitur lainnya memiliki nilai mendekati nol, yang mengindikasikan bahwa kontribusi mereka terhadap model sangat kecil atau bahkan dapat diabaikan. Dengan demikian, fitur-fitur dengan skor tinggi perlu mendapat perhatian lebih dalam analisis atau pengembangan model, sedangkan fitur-fitur dengan skor rendah bisa dipertimbangkan untuk dihapus guna menyederhanakan model dan mengurangi kompleksitas.

3.1.3 Data Preparation

Tahap data preparation adalah proses awal dalam analisis data yang meliputi serangkaian langkah untuk mengolah, membersihkan, dan mempersiapkan data agar siap digunakan dalam model analisis. Proses ini penting dalam memastikan bahwa data yang digunakan berkualitas tinggi dan relevan untuk tujuan analisis. Berikut ini tahapan Data Preparation yang kami lakukan dalam mengolah dataset.

3.1.3.3 Data Selection

Data selection adalah proses pemilihan data yang relevan untuk digunakan dalam pengerjaan proyek. Pada proyek kali ini data yang digunakan berasal dari dataset yang sudah ditentukan terlebih dahulu. Oleh karena itu dilakukan terlebih dahulu menampilkan ringkasan dataset dan memeriksa struktur dataset. Hasil data selection dapat dilihat dari Gambar 3 berikut

```
In [23]: # Replace 'data' with either 'train_data' or 'test_data',
# depending on which dataset's information you want to view.

# To view train data information:
train_data.info()

# To view test data information:
# test_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Id                   1460 non-null   int64
1   MSSubClass           1460 non-null   int64
2   MSZoning             1460 non-null   object
3   LotFrontage         1201 non-null   float64
4   LotArea             1460 non-null   int64
5   Street              1460 non-null   object
6   Alley               91 non-null     object
7   LotShape            1460 non-null   object
8   LandContour         1460 non-null   object
9   Utilities           1460 non-null   object
10  LotConfig           1460 non-null   object
11  LandSlope           1460 non-null   object
12  Neighborhood        1460 non-null   object
13  Condition1          1460 non-null   object
14  Condition2          1460 non-null   object
15  BldgType            1460 non-null   object
16  HouseStyle          1460 non-null   object
17  OverallQual         1460 non-null   int64
18  OverallCond         1460 non-null   int64
19  YearBuilt            1460 non-null   int64
20  YearRemodAdd        1460 non-null   int64
21  RoofStyle           1460 non-null   object
22  RoofMatl            1460 non-null   object
23  Exterior1st         1460 non-null   object
24  Exterior2nd         1460 non-null   object
25  MasVnrType          588 non-null     object
26  MasVnrArea          1452 non-null   float64
27  ExterQual            1460 non-null   object
28  ExterCond           1460 non-null   object
29  Foundation          1460 non-null   object
30  BsmtQual            1423 non-null   object
```

Gambar 3. Data Selection

3.1.3.4 Data Cleaning

Pada proyek ini, data cleaning tidak perlu dilakukan, karena data yang diberikan sudah bersih. Gambar 4 menunjukkan proses identifikasi missing values pada variabel numerik dalam dataset sebagai bagian dari langkah data cleaning. Proses ini dilakukan dengan mengecek nilai yang hilang menggunakan fungsi `isnull()` dan menghitung proporsinya dengan `mean()`. Hasilnya menunjukkan bahwa dari 40 fitur numerik, hanya 3 fitur yang memiliki missing values, yaitu *LotFrontage* dengan proporsi tertinggi sebesar 17.73%, diikuti oleh *GarageYrBlt* sebesar 6.53% dan *MasVnrArea* sebesar 0.46%. Sementara itu, fitur lainnya memiliki proporsi 0.0, yang berarti tidak ada data yang hilang. Langkah ini penting untuk memahami sejauh mana data numerik memiliki nilai yang hilang sehingga dapat ditangani dengan metode yang tepat, seperti imputasi atau penghapusan data, guna memastikan dataset bersih dan siap digunakan untuk proses analisis atau pemodelan lebih lanjut. Pencarian missing value pada file `fktpkapitasi.dta` dapat dilihat pada gambar 4.

```
In [30]: # Missing values in our numerical variables
Num_V = X_train[vars_num].isnull().mean().sort_values(ascending=False)
print(Num_V)
len(Num_V)
```

LotFrontage	0.177321
GarageYrBlt	0.056317
MasVnrArea	0.004566
WoodDeckSF	0.000000
BedroomAbvGr	0.000000
KitchenAbvGr	0.000000
TotRmsAbvGrd	0.000000
Fireplaces	0.000000
GarageCars	0.000000
GarageArea	0.000000
MSSubClass	0.000000
HalfBath	0.000000
EnclosedPorch	0.000000
3SsnPorch	0.000000
ScreenPorch	0.000000
PoolArea	0.000000
MiscVal	0.000000
MoSold	0.000000
OpenPorchSF	0.000000
FullBath	0.000000
BsmtHalfBath	0.000000
BsmtFullBath	0.000000
GrLivArea	0.000000
LowQualFinSF	0.000000
2ndFlrSF	0.000000
1stFlrSF	0.000000
TotalBsmtSF	0.000000
BsmtUnfSF	0.000000
BsmtFinSF2	0.000000
BsmtFinSF1	0.000000
YearRemodAdd	0.000000
YearBuilt	0.000000
OverallCond	0.000000
OverallQual	0.000000
LotArea	0.000000
YrSold	0.000000

dtype: float64

Gambar 4. Missing value pada file `fktpkapitasi.dta`

3.1.3.5 Data Tranformation

Subbab ini menjelaskan transformasi data yang dilakukan dengan memahami struktur dataset melalui fungsi `train_data.info()`, yang menampilkan informasi tentang 81 atribut, termasuk tipe data (`int64`, `float64`, `object`) dan kelengkapan nilai. Dataset memiliki 1460 baris, dengan beberapa atribut tidak lengkap, seperti `LotFrontage` (1201 nilai) dan `Alley` (91 nilai). Transformasi dilakukan dengan memilih atribut relevan, seperti `OverallQual`, `YearBuilt`, dan `SalePrice` untuk analisis, menangani nilai yang hilang pada atribut seperti `Alley`, serta menyesuaikan tipe data agar sesuai untuk analisis lebih lanjut. Transformasi ini memastikan dataset bersih, lengkap, dan relevan untuk mendukung tujuan proyek, dapat dilihat pada gambar 5.

```
In [23]: # Replace 'data' with either 'train_data' or 'test_data',
#         # depending on which dataset's information you want to view.

# To view train_data information:
train_data.info()

# To view test data information:
# test_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Id                   1460 non-null  int64
1   MSSubClass           1460 non-null  int64
2   MSZoning             1460 non-null  object
3   LotFrontage         1201 non-null  float64
4   LotArea             1460 non-null  int64
5   Street              1460 non-null  object
6   Alley               91 non-null    object
7   LotShape            1460 non-null  object
8   LandContour         1460 non-null  object
9   Utilities           1460 non-null  object
10  LotConfig           1460 non-null  object
11  LandSlope           1460 non-null  object
12  Neighborhood         1460 non-null  object
13  Condition1          1460 non-null  object
14  Condition2          1460 non-null  object
15  BldgType            1460 non-null  object
16  HouseStyle          1460 non-null  object
17  OverallQual         1460 non-null  int64
18  OverallCond         1460 non-null  int64
19  YearBuilt           1460 non-null  int64
20  YearRemodAdd        1460 non-null  int64
21  RoofStyle           1460 non-null  object
22  RoofMatl            1460 non-null  object
23  Exterior1st         1460 non-null  object
24  Exterior2nd         1460 non-null  object
25  MasVnrType          588 non-null   object
26  MasVnrArea          1452 non-null  float64
27  ExterQual            1460 non-null  object
28  ExterCond           1460 non-null  object
29  Foundation          1460 non-null  object
```

Gambar 5. Data Transformation

3.1.3.6 Feature Engineering

Kami melakukan Feature engineering untuk menyederhanakan representasi data dengan menggabungkan kolom terkait. Tujuannya adalah mengurangi redundansi, meningkatkan relevansi fitur. Berikut tahapan dalam melakukan Feature Engineering.

1. Membuat Fitur Baru: "Date"

Fitur **Date** menggabungkan informasi terkait tanggal yang dapat mempengaruhi harga rumah, seperti tahun dibangun (`YearBuilt`) dan tahun renovasi terakhir (`YearRemodAdd`). Fitur ini mengukur waktu yang telah berlalu sejak rumah dibangun atau direnovasi, dan dapat memberikan wawasan tambahan untuk prediksi harga.

Formula:

- **Date** = Tahun sekarang - **YearBuilt**
- **Date** = Tahun sekarang - **YearRemodAdd** Fitur ini memberikan informasi mengenai umur rumah dan apakah rumah tersebut telah direnovasi baru-baru ini.

2. Membuat Fitur Baru: “Numerik Features”

Menambahkan beberapa fitur numerik baru untuk memperkaya data yang dapat digunakan dalam model prediksi. Beberapa fitur numerik baru yang bisa ditambahkan adalah:

- **Price per Square Foot:**
Menghitung harga per meter persegi (atau per kaki persegi) dengan membagi harga jual rumah dengan total luas rumah.
Formula:
$$\text{Price per Square Foot} = \frac{\text{SalePrice}}{\text{Total Square Footage}}$$

$$\text{Price per Square Foot} = \text{Total Square Footage} \div \text{SalePrice}$$
- **Price per Bedroom:**
Menghitung harga jual per kamar tidur. Ini bisa membantu untuk melihat apakah harga rumah berkorelasi dengan jumlah kamar tidur.
Formula:
$$\text{Price per Bedroom} = \frac{\text{SalePrice}}{\text{TotRmsAbvGrd}}$$

$$\text{Price per Bedroom} = \text{TotRmsAbvGrd} \div \text{SalePrice}$$

(dimana `TotRmsAbvGrd` adalah jumlah total kamar di atas tanah).
- **Garage Price:**
Menghitung harga jual per area garasi untuk memberikan wawasan tambahan tentang seberapa besar nilai garasi dalam harga rumah.
Formula:
$$\text{Garage Price} = \frac{\text{SalePrice}}{\text{GarageArea}}$$

$$\text{Garage Price} = \text{GarageArea} \div \text{SalePrice}$$

3. Membuat Fitur Baru: “MiscFeature”

Fitur **MiscFeature** mencakup informasi tambahan yang tidak dapat digolongkan dalam kategori tertentu, seperti adanya fitur khusus di rumah yang dapat mempengaruhi harga. Kolom ini dapat digabungkan dari berbagai jenis fitur tambahan yang ada, seperti:

- **PoolQC:** Kualitas kolam renang, apakah ada atau tidak.
- **Fence:** Jenis pagar rumah, apakah ada atau tidak.
- **MiscFeature:** Fitur tambahan lainnya seperti sistem pemanas, atau adanya dek yang belum terhitung dalam fitur lainnya.

Untuk menciptakan fitur ini, kita bisa mengubah status kolom yang ada menjadi biner (0 atau 1) untuk menunjukkan apakah fitur tersebut ada pada rumah atau tidak. Misalnya:

- **HasPool** = 1 jika rumah memiliki kolam renang, 0 jika tidak.
- **HasFence** = 1 jika rumah memiliki pagar, 0 jika tidak.

3.1.4 Modelling

Pada tahap modelling kami menggunakan 3 model dalam mengerjakan proyek ini, yaitu 2 model sebagai pembandingan dan 1 sebagai model utama. Dengan Random Forest dan Neural Network sebagai model pembandingan dan XGBoost sebagai model utama.

3.1.4.3 *Random Forest*

Pada proyek ini, algoritma *Random Forest* dipilih karena kemampuannya yang baik dalam menangani data beragam dan meminimalkan risiko *overfitting*.

Langkah-langkah implementasi meliputi:

1. Membagi Data menjadi Train dan Test

Membagi dataset menjadi data pelatihan dan pengujian agar dapat mengevaluasi performa model secara akurat dengan proporsi 80% untuk train dan 20% untuk test.

```
X_train_tfr=price_pipe.transform(X_train) # Transformation for train se
X_test_tfr=price_pipe.transform(X_test) # Transformation for test set
```

3.1.4.4 *XGBoost*

Pada proyek ini, algoritma XGBoost digunakan dipilih karena performanya yang unggul pada data yang memiliki korelasi fitur yang kompleks.

Langkah-langkah implementasi meliputi:

1. Pembuatan Model Xgboost Regression Model

Dapat kita lihat pada gambar 6.

```
In [57]: # XG-Boost Regressor
# Create an xgboost regression model

model = xgb.XGBRegressor(n_estimators=100, max_depth=7, eta=0.1, subsample=0.7, colsample_by
```

Gambar 6. Data Split XGBoost

2. Menentukan Hyperparameter untuk Pencarian Optimal

Melakukan pencarian hyperparameter terbaik untuk model XGBoost agar kinerjanya optimal. Dapat kita lihat pada gambar 7.


```
In [ ]: # Train on training set
        model.fit(X_train_tfr, y_train)

Out[ ]: XGBRegressor(base_score=None, booster=None, callbacks=None,
                    colsample_bylevel=None, colsample_bynode=None,
                    colsample_bytrees=0.8, device=None, early_stopping_rounds=None,
                    enable_categorical=False, eta=0.1, eval_metric=None,
                    feature_types=None, gamma=None, grow_policy=None,
                    importance_type=None, interaction_constraints=None,
                    learning_rate=None, max_bin=None, max_cat_threshold=None,
                    max_cat_to_onehot=None, max_delta_step=None, max_depth=7,
                    max_leaves=None, min_child_weight=None, missing=nan,
                    monotone_constraints=None, multi_strategy=None, n_estimators=100,
                    n_jobs=None, num_parallel_tree=None, ...)
```

Gambar 7. Data Split XGBoost

3. Melakukan Prediksi

Menguji performa model dengan melakukan prediksi pada data uji.

```
pred = model.predict(X_train_tfr)
```

3.1.5 Evaluation

Tahap Evaluasi adalah proses menilai kinerja model setelah model dilatih menggunakan data pelatihan. Tahap ini penting untuk memastikan model dapat melakukan generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Tujuan dari tahap ini untuk mengukur seberapa baik model dalam membuat prediksi pada data uji. Memastikan model tidak hanya menyesuaikan diri dengan data pelatihan, tetapi juga dapat bekerja dengan baik pada data baru.

Metode Evaluasi yang kami gunakan antara lain:

- **Akurasi:** Proporsi jumlah prediksi yang benar dibandingkan dengan total prediksi.
- **Precision:** Mengukur proporsi prediksi benar dari total prediksi positif (benar positif / (benar positif + salah positif)). Ini penting untuk menilai seberapa banyak dari prediksi positif yang benar-benar positif.
- **Recall:** Mengukur proporsi dari benar positif yang berhasil ditemukan oleh model (benar positif / (benar positif + salah negatif)). Ini menunjukkan seberapa sensitif model dalam mendeteksi positif.
- **F1-Score:** Rata-rata antara precision dan recall, memberikan gambaran lebih lengkap tentang model terutama ketika ada ketidakseimbangan kelas.
- **Support:** Jumlah instance yang ada di masing-masing kelas dalam dataset.

3.1.5.3 Evaluasi Model XGboost

```
In [84]: # Evaluate the model:
# Evaluate performance using the mean squared error and the root of the mean squared error
pred = model_no_pro.predict(X_train)
print('linear train mse: {}'.format(mean_squared_error(y_train, pred)))
print('linear train rmse: {}'.format(sqrt(mean_squared_error(y_train, pred))))
print()
pred = model_no_pro.predict(X_test)
print('linear test mse: {}'.format(mean_squared_error(y_test, pred)))
print('linear test rmse: {}'.format(sqrt(mean_squared_error(y_test, pred))))

linear train mse: 30921646.87456972
linear train rmse: 5560.723592714326

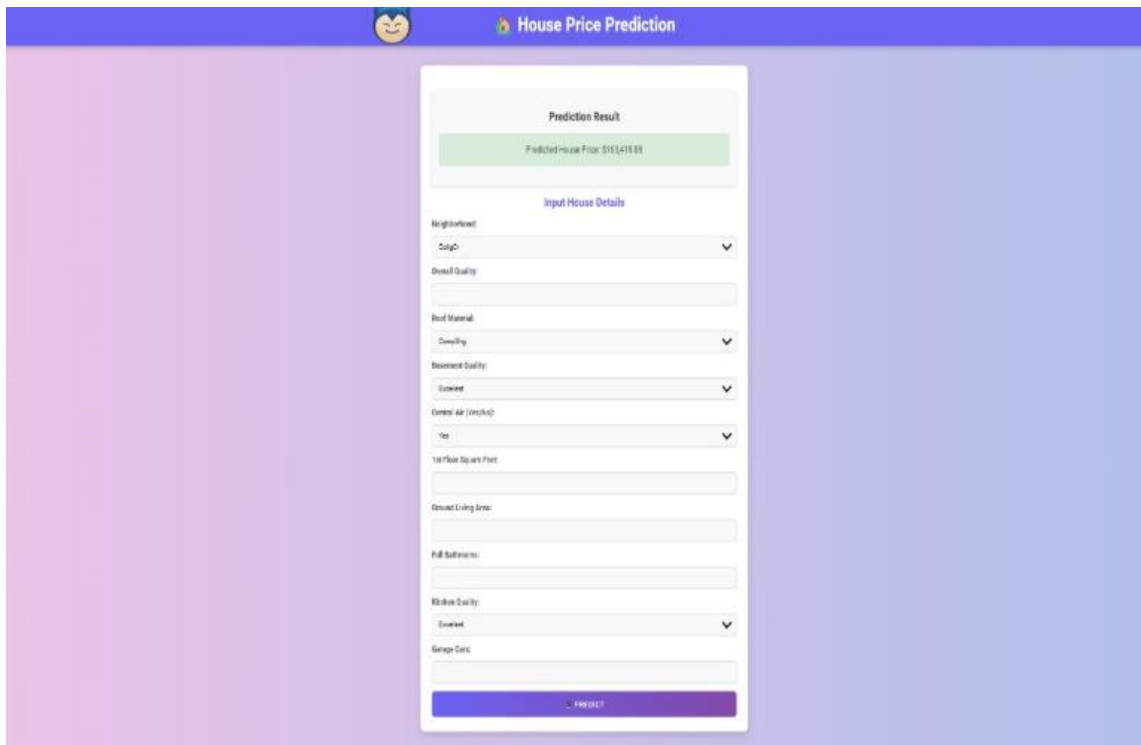
linear test mse: 987064343.7865986
linear test rmse: 31417.580170767425
```

Gambar 8. Evaluasi Model Random Forest

Dalam gambar yang Anda unggah, terlihat bahwa proses evaluasi model dilakukan dengan menghitung **Mean Squared Error (MSE)** dan **Root Mean Squared Error (RMSE)** pada data pelatihan (training) dan data uji (test). Hasil evaluasi menunjukkan nilai MSE dan RMSE yang cukup besar, terutama pada data uji, yang mengindikasikan bahwa model yang digunakan belum berkinerja dengan baik. Nilai RMSE yang tinggi, baik pada data pelatihan maupun uji, menunjukkan bahwa prediksi model jauh dari nilai sebenarnya. Kami menghadapi kesulitan dalam evaluasi model karena meskipun nilai akurasi yang didapatkan cukup tinggi pada data pelatihan, hasil pada data uji menunjukkan kesalahan yang besar. Hal ini menunjukkan bahwa model mungkin mengalami **overfitting**, di mana model terlalu menyesuaikan diri dengan data pelatihan dan tidak dapat menggeneralisasi dengan baik pada data uji. Akibatnya, meskipun akurasi pada data pelatihan terlihat baik, akurasi yang buruk pada data uji memperburuk kinerja model secara keseluruhan, membuat evaluasi model menjadi tidak optimal.

3.1.6 Deployment

Pada tahap deployment adalah langkah di mana model yang telah dikembangkan dan dilatih siap untuk digunakan. Dalam konteks aplikasi yang kami buat adalah untuk memprediksi orang depresi atau tidak. Dalam tahap deployment ini kami menggunakan framework untuk mengembangkan model yang kami buat. Berikut adalah hasil deployment yang telah kami buat, dapat dilihat pada Gambar 9.

The screenshot shows a web application titled "House Price Prediction" with a purple header. The main content area has a light purple background. In the center, there is a white card. At the top of the card, under the heading "Prediction Result", there is a green box displaying "Predicted House Price: \$15,418.88". Below this, under the heading "Input House Details", there is a form with several input fields: "Neighborhood" (a dropdown menu showing "CollgCr"), "Overall Quality" (a text input field), "Roof Material" (a dropdown menu showing "CemGr", with "Shingle" visible below it), "Basement Quality" (a dropdown menu showing "Good", with "Average" visible below it), "Central Air (yes/no)" (a dropdown menu showing "Yes", with "No" visible below it), "Total Floor Square Feet" (a text input field), "Overall Living Area:" (a text input field), "Full Bathrooms:" (a text input field), "Kitchen Quality" (a dropdown menu showing "Good", with "Average" visible below it), and "Garage Cars" (a text input field). At the bottom of the card is a purple button labeled "PREDICT".

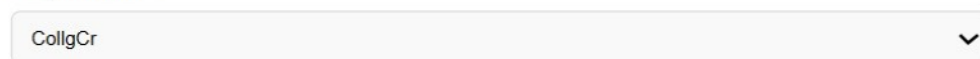
Gambar 9. Deployment

Untuk detail fitur-fitur yang kami buat dapat dilihat di bawah ini:

1. Neighborhood

Pada bagian ini diberikan pilihan yang menjelaskan kondisi dari lingkungan sekitar rumah

Neighborhood:

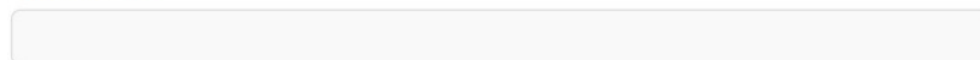
The screenshot shows a dropdown menu for the "Neighborhood" field. The selected option is "CollgCr". There is a downward arrow icon on the right side of the dropdown.

Gambar 10. Neighborhood

2. Overall Quality

Pada bagian ini memberikan informasi dari kualitas dari keseluruhan rumah tersebut

Overall Quality:

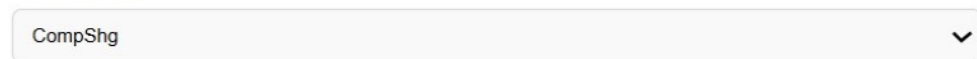
The screenshot shows a text input field for the "Overall Quality" field. The field is currently empty.

Gambar 11. Overall Quality

3. Roof Material

Pada bagian ini diberikan pilihan untuk menjelaskan material yang digunakan untuk bagian atap

Roof Material:

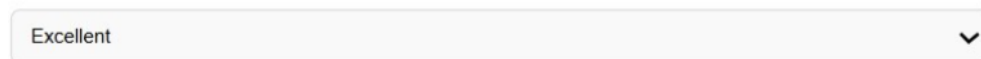
A light gray rectangular dropdown menu with rounded corners. It contains the text 'CompShg' on the left and a small downward-pointing chevron icon on the right.

Gambar 12. Roof Material

4. Basement Quality

Pada bagian ini diberikan penjelasan kualitas dari basement rumah yang akan di jual

Basement Quality:

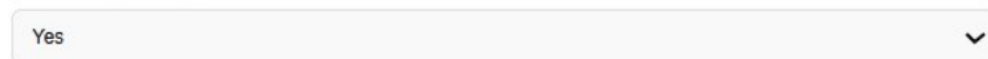
A light gray rectangular dropdown menu with rounded corners. It contains the text 'Excellent' on the left and a small downward-pointing chevron icon on the right.

Gambar 13. Basement Quality

5. Central Air

Pada bagian ini memberikan informasi apakah rumah tersebut memiliki alat sirkulasi udara atau tidak.

Central Air (Yes/No):

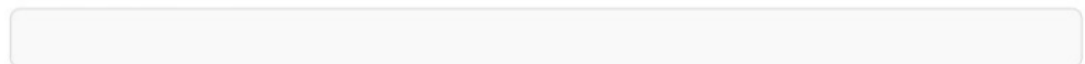
A light gray rectangular dropdown menu with rounded corners. It contains the text 'Yes' on the left and a small downward-pointing chevron icon on the right.

Gambar 14. Central Air

6. 1st Floor Square Feet

Pada bagian ini memberikan informasi tentang luas dari lantai satu rumah tersebut.

1st Floor Square Feet:

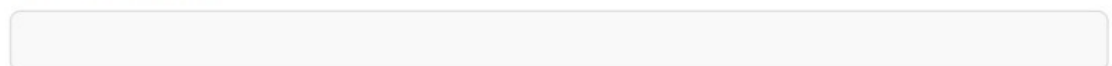
A light gray rectangular text input field with rounded corners and a thin border.

Gambar 15. 1st Floor Square Feet

7. Ground Living Area

Pada bagian ini memberikan informasi tentang luas tanah dari rumah tersebut.

Ground Living Area:

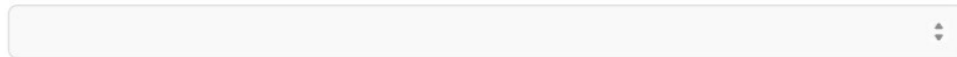
A light gray rectangular text input field with rounded corners and a thin border.

Gambar 16. Ground Living Area

8. Full Bathrooms

Pada bagian ini memberikan informasi tentang jumlah kamar mandi yang ada didalam rumah tersebut.

Full Bathrooms:

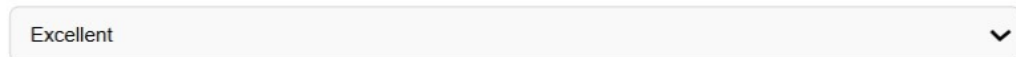


Gambar 17. Full Bathrooms

9. Kitchen Quality

Pada bagian ini memberikan informasi tentang kualitas dari dapur rumah tersebut.

Kitchen Quality:

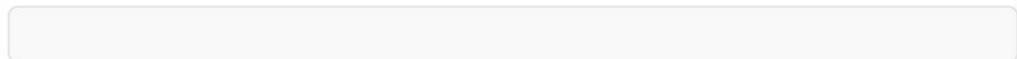


Gambar 18. Kitchen Quality

10. Garage Cars

Pada bagian ini memberikan informasi jumlah garasi yang ada pada rumah tersebut.

Garage Cars:



Gambar 19. Garage Cars

3.2 Timeline

Berikut ini adalah Timeline dari pengerjaan proyek yang akan kami lakukan sampai dengan selesai:

Aktivitas	Sub Aktivitas	Detail	Week																																
			12					13					14					15					16												
			November														Desember																		
			11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13
Persiapan	Pemilihan Kasus dan	Pemilihan Kasus																																	
		Penentuan Algoritma																																	
Pelaksanaan	Business Understanding	Menentukan Objektif Bisnis																																	
		Menentukan Tujuan Bisnis																																	
		Membuat Rencana Proyek																																	
	Data Understanding	Mengumpulkan Data																																	
		Memvalidasi Data																																	
	Data Preparation	Memilih Data																																	
		Membersihkan Data																																	
		Mengkonstruksi Data																																	
		Menentukan Label Data																																	
	Modeling	Membangun Skema Pengujian																																	
		Membangun Model																																	
	Model Evaluation	Mengevaluasi Hasil Pemodelan																																	
		Melakukan Review Proses Pemodelan																																	
Deployment	Melakukan Deployment Model																																		
	Membuat laporan akhir Proyek																																		

Gambar 44. Timeline

BAB IV

KESIMPULAN

4.1 Kesimpulan

Kesimpulan yang kami dapat dari pengerjaan proyek ini adalah:

1. Kami mengalami peningkatan yang signifikan dalam keterampilan dan pengetahuan di bidang data mining. Selama pengerjaan proyek ini, kami berhasil menyelesaikan proyek praktis yang memungkinkan penerapan teori yang telah dipelajari ke dalam implementasi nyata.
2. Model XGBoost yang digunakan dalam prediksi harga rumah mampu menghasilkan hasil yang lebih baik dibandingkan dengan beberapa model lain. Namun, tingkat akurasi model masih tergolong rendah, sementara nilai RMSE (Root Mean Square Error) dan MSE (Mean Square Error) masih cukup tinggi.
3. Model dapat melakukan prediksi sesuai dengan data input yang dimasukkan, namun hasil prediksinya masih perlu ditingkatkan untuk memastikan keandalan dalam aplikasi nyata.
4. Proses deployment berjalan dengan baik, memungkinkan model untuk digunakan secara langsung melalui website. Namun, performa model yang masih perlu pengembangan menjadi fokus utama dalam iterasi berikutnya.

Secara keseluruhan, pengalaman selama mengerjakan proyek ini sangat bermanfaat dan memberikan fondasi yang kuat dalam pengembangan keterampilan kami di bidang data mining. Kami menyadari bahwa pengembangan lebih lanjut diperlukan, terutama dalam meningkatkan akurasi model dan menurunkan nilai RMSE serta MSE. Kami optimis dapat menghadapi tantangan ini dan berkontribusi secara lebih signifikan dalam proyek-proyek analisis data di masa depan.

BAB V

PEMBAGIAN PEKERJAAN

1	Mikhael Janugrah Pakpahan	Deployment
2	Bobby Willy Siagian	Data Understanding & Data Preparation
3	Aldi Jeremy Simamora	Deployment, Modelling

REFERENSI

1. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv preprint arXiv:1603.02754*. <https://arxiv.org/abs/1603.02754>
2. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.
3. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
4. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.
5. Kaggle. (n.d.). House Prices: Advanced Regression Techniques. Retrieved from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

LAMPIRAN