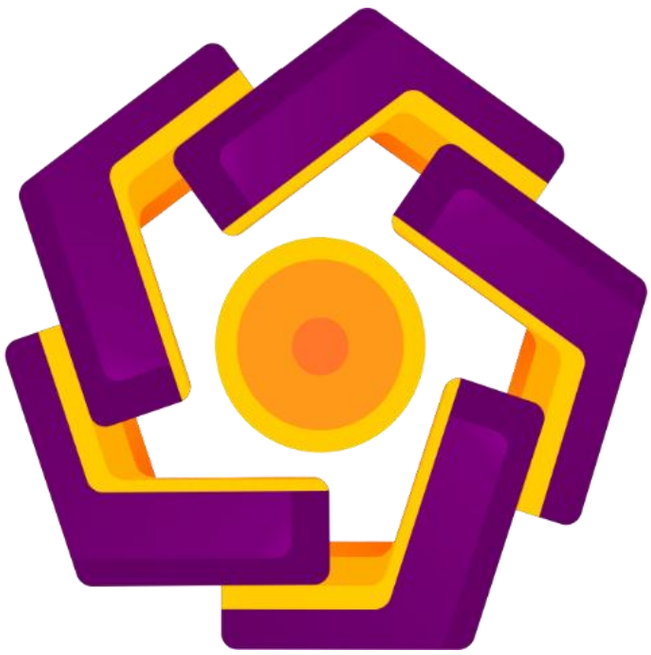
UJIAN AKHIR SEMESTER

BIG DATA & PREDICTIVE ANALYTICS LANJUT

(Housing Price)



DOSEN PENGAMPU:

Ajie Kusuma Wardhana

disusun oleh:

Rofi Almahendra A	**22.11.5067**
Nurezky Agam	**22.11.5054**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM

YOGYAKARTAYOGYAKARTA

2025

Soal Ujian *(disesuaikan dengan sifat ujian)*

1. Berdasarkan apa yang sudah Anda pelajari, silahkan gunakan kemampuan anda untuk menyelesaikan sebuah menggunakan *classification* yang melibatkan penggunaan Machine Learning. (SCPMK 1534113, 50 Poin)

   a Pilih satu bidang yang Anda beserta rekan tim minati, jabarkan alasan pemilihan bidang tersebut dan jelaskan apa yang ingin dicapai dengan memilih topik ini.

   Saya memilih klasifikasi popular anime karena anime merupakan salah satu bentuk seni dan hiburan yang memiliki basis penggemar yang sangat luas di seluruh dunia. Dalam dunia yang semakin terhubung, pemahaman tentang faktor-faktor yang memengaruhi popularitas anime dapat membantu berbagai pihak, seperti platform streaming, studio produksi, dan komunitas penggemar, untuk menyajikan konten yang lebih relevan dan sesuai dengan preferensi audiens.

   Apa yang Ingin Dicapai :

   - Identifikasi Faktor Penentu Popularitas Menggunakan teknik machine learning untuk memahami hubungan antara berbagai fitur (seperti skor, jumlah episode, genre, dan jumlah penonton) terhadap popularitas suatu anime.
   - Peningkatan Rekomendasi Konten Dengan memahami tren popularitas, kami ingin membantu platform streaming atau pengembang aplikasi untuk memberikan rekomendasi anime yang lebih sesuai dengan minat pengguna.
   - Dukungan pada Produksi Konten Memberikan wawasan kepada studio anime tentang elemen-elemen yang perlu dipertimbangkan dalam menciptakan anime yang lebih diminati.
   - Peningkatan Pengalaman Pengguna Membantu penggemar menemukan anime yang sesuai dengan preferensi mereka, berdasarkan pola popularitas yang dianalisis.

   b Ceritakan proses mendapatkan data dan informasi lengkap mengenai data tersebut (seperti waktu, penjelasan setiap kolom, sumber dll). Data yang digunakan harus data terbaru dengan range 1-4 tahun kebelakang.

   Pengambilan dataset dilakukan pada 17/1/2025.

   Dataset ini memiliki 13 kolom:

   1. price
   2. area
   3. bedroom
   4. bathroom
   5. Stories
   6. mainroad
   7. guestroom
   8. basement
   9. Hot waterheating

10. airconditioning
11. Parking
12. Prefarea
13. Furnishing status

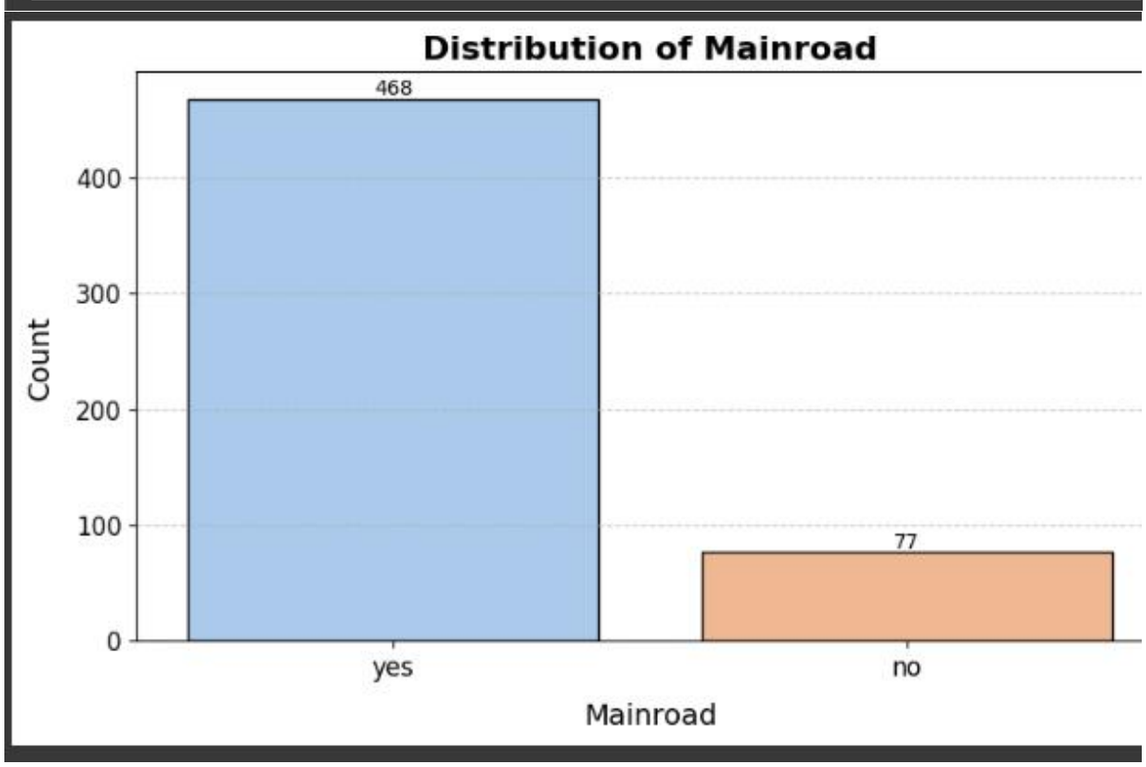
Link Dataset : https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction/code

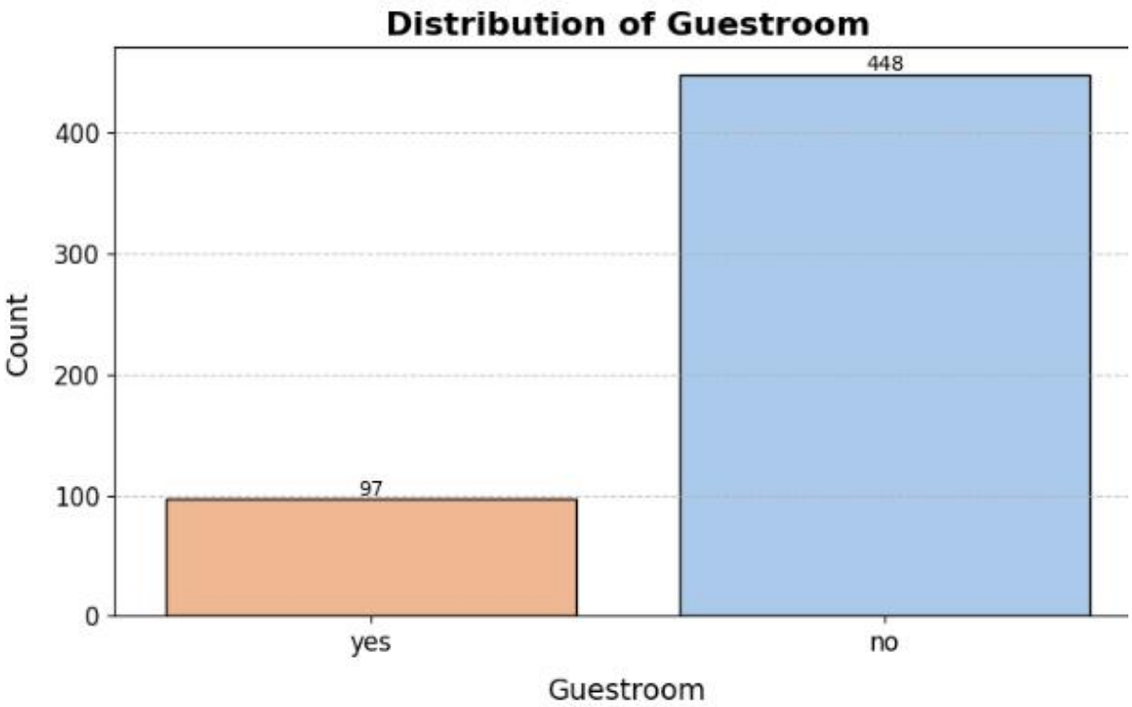
c Lakukan pre-processing data dengan memeriksa tipe data, mengganti nama kolom, memeriksa nilai null, mengubah tipe data (agar bisa di proses), menampilkan summary, dan menampilkan matriks korelasinya menggunakan metode metode yang pernah dipelajari.

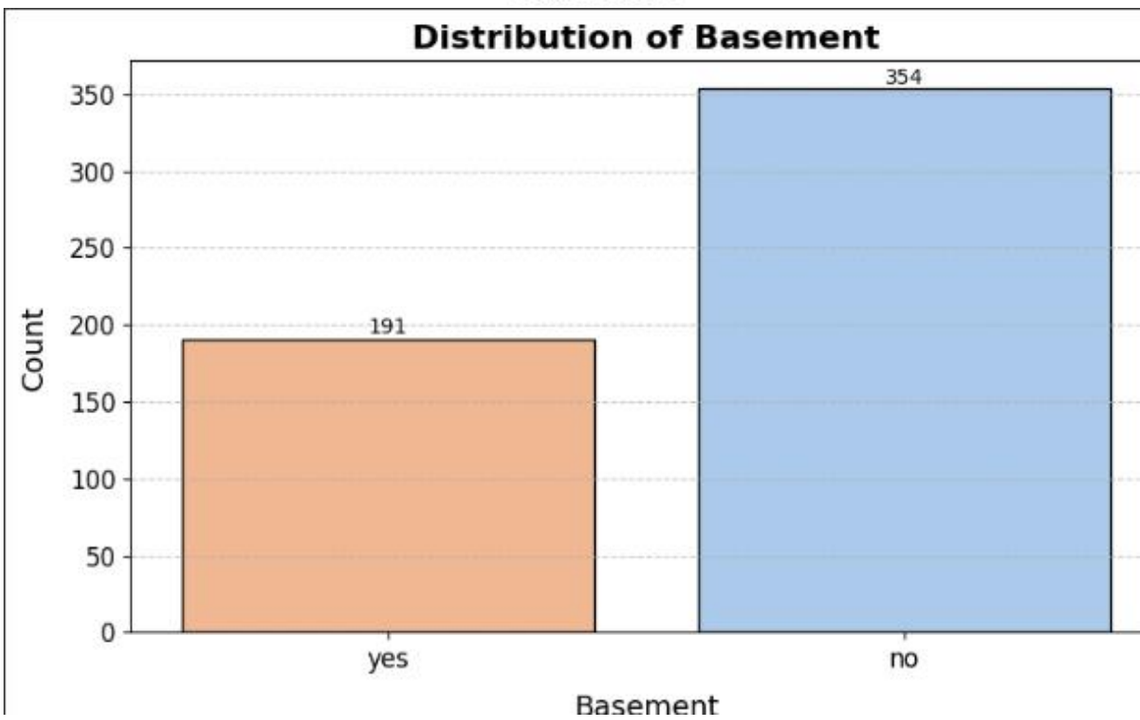
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   price             545 non-null    int64 
 1   area              545 non-null    int64 
 2   bedrooms          545 non-null    int64 
 3   bathrooms         545 non-null    int64 
 4   stories           545 non-null    int64 
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64 
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
memory usage: 55.5+ KB
```

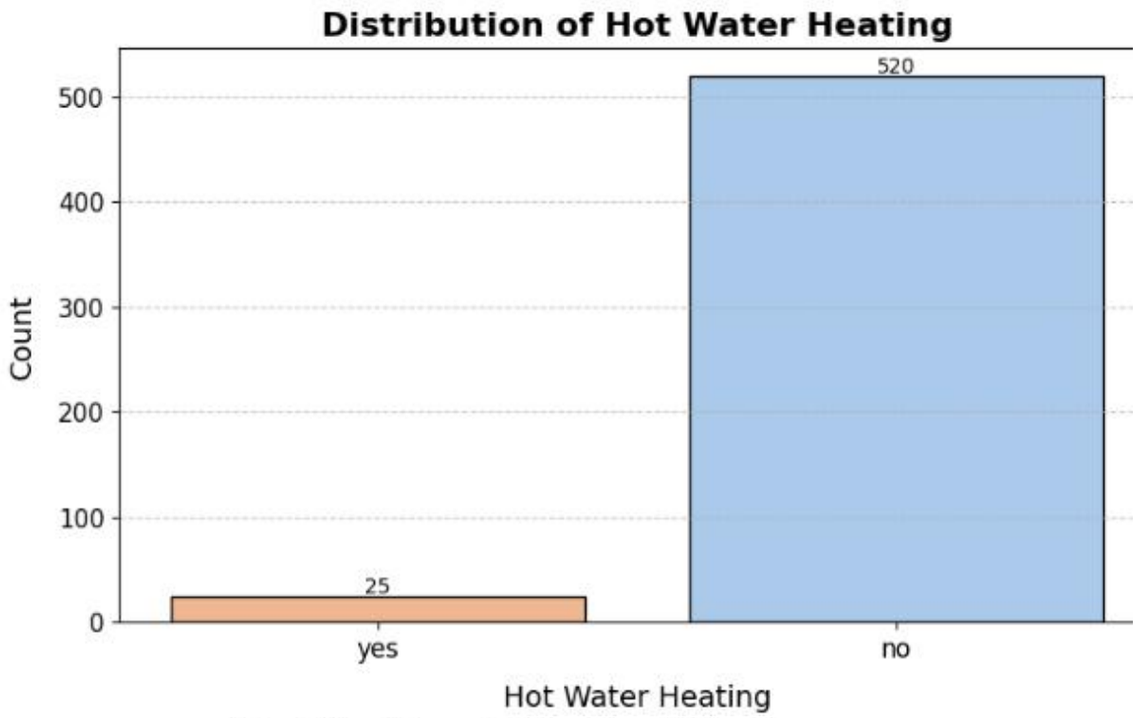
	price	area	bedrooms	bathrooms	stories	parking
count	5.450000e+02	545.000000	545.000000	545.000000	545.000000	545.000000
mean	4.766729e+06	5150.541284	2.965138	1.286239	1.805505	0.693578
std	1.870440e+06	2170.141023	0.738064	0.502470	0.867492	0.861586
min	1.750000e+06	1650.000000	1.000000	1.000000	1.000000	0.000000
25%	3.430000e+06	3600.000000	2.000000	1.000000	1.000000	0.000000
50%	4.340000e+06	4600.000000	3.000000	1.000000	2.000000	0.000000
75%	5.740000e+06	6360.000000	3.000000	2.000000	2.000000	1.000000
max	1.330000e+07	16200.000000	6.000000	4.000000	4.000000	3.000000

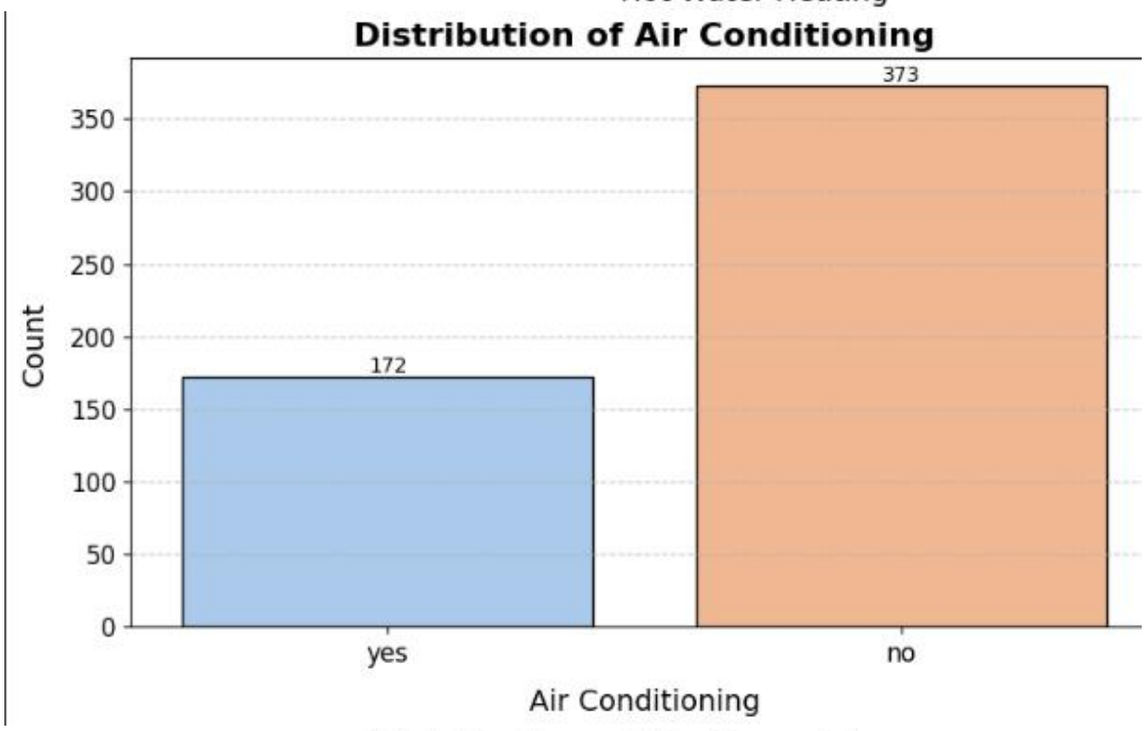
d Gunakan exploratory data analysis (EDA) untuk melihat sudut padang yang ada mengenai data (minimal 4) dua diantaranya bar dan pie chart, 2 diantaranya bebas. Berikan penjelasan.

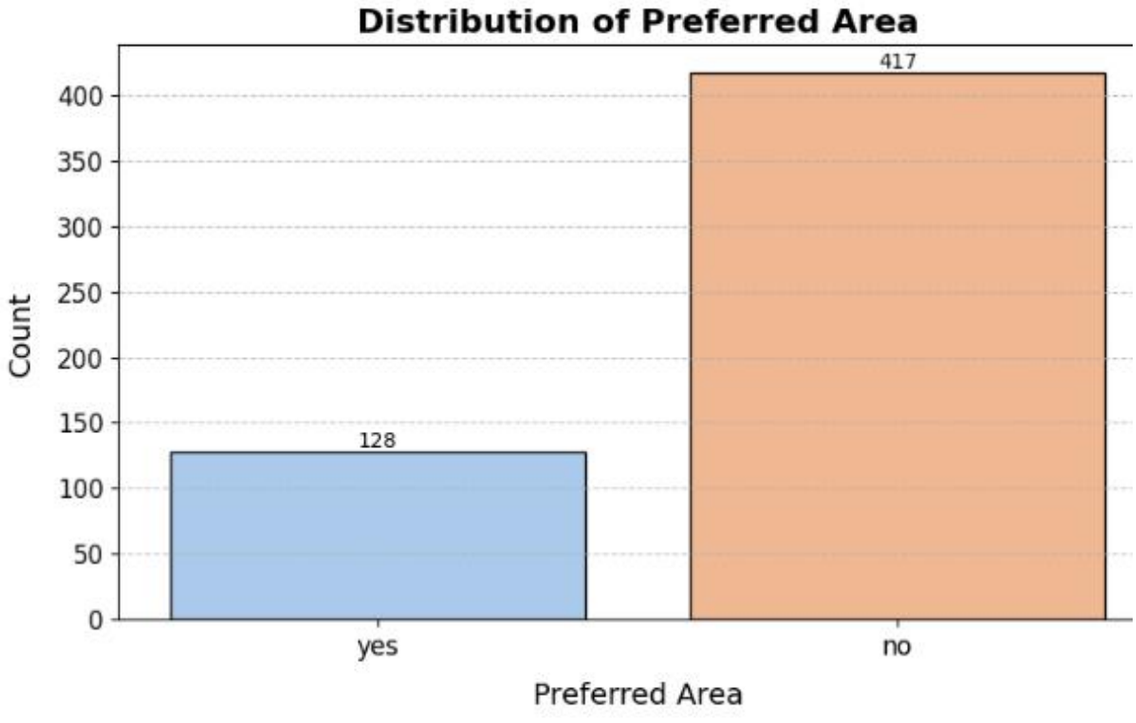


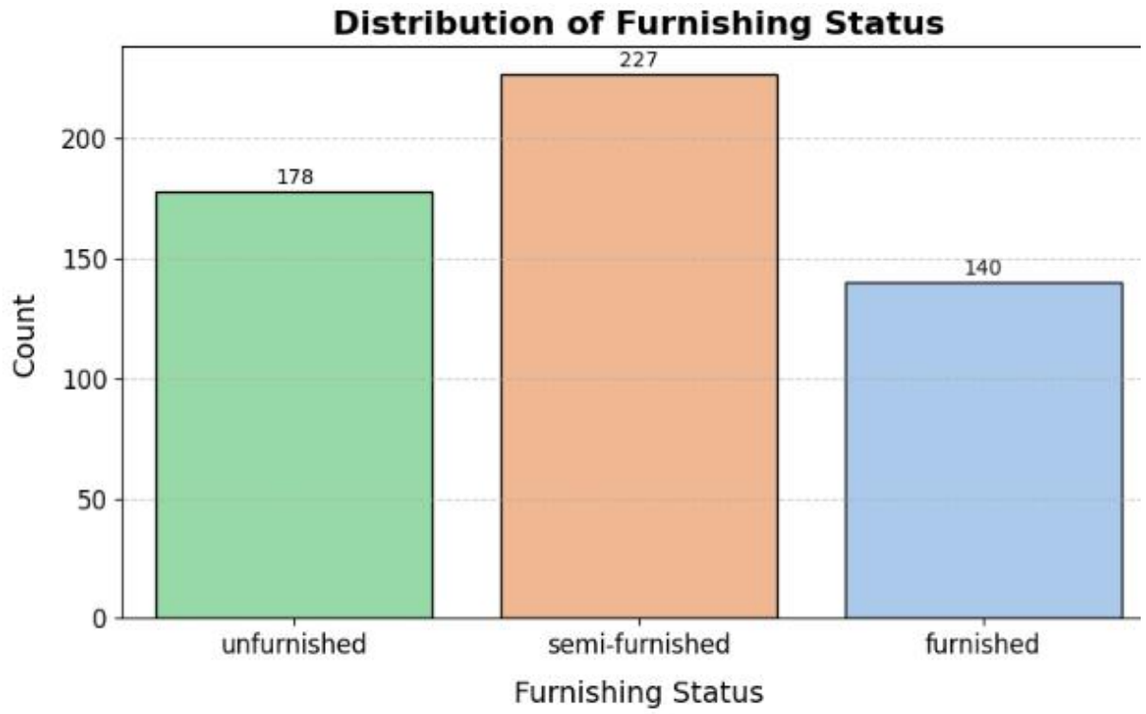

Distribution of Guestroom
448
97
400
300
200
100
0
Count
yes
no
Guestroom

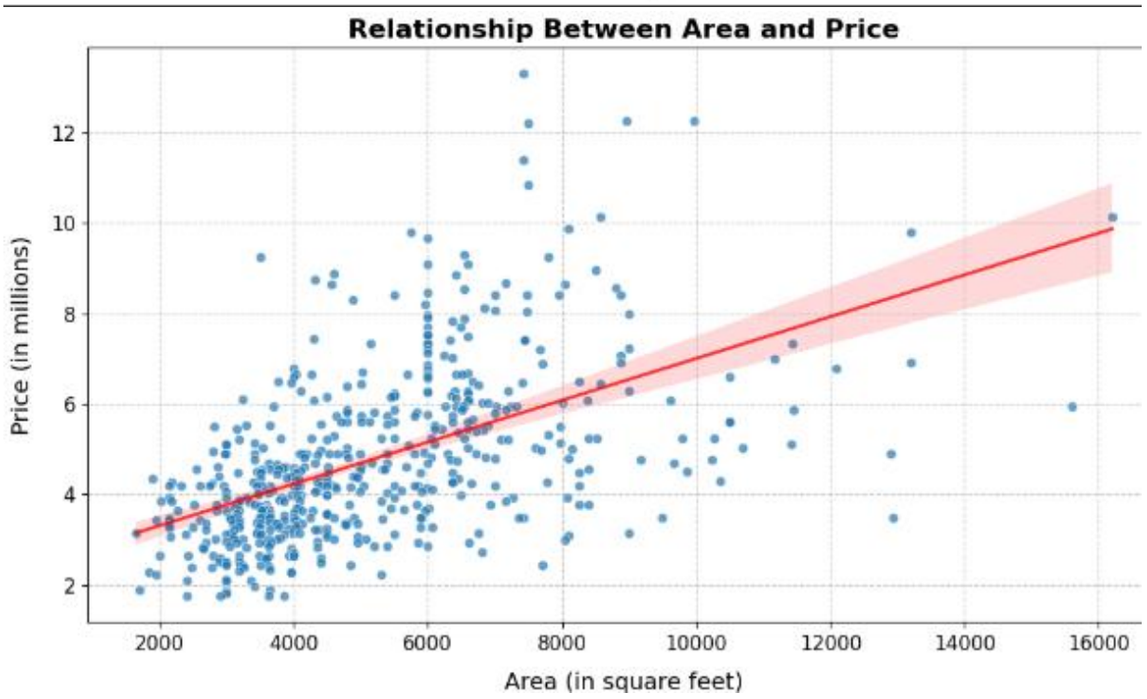

Distribution of Basement
354
191
350
300
250
200
150
100
50
0
Count
yes
no
Basement


Distribution of Hot Water Heating
Count
520
25
0
100
200
300
400
500
yes
no
Hot Water Heating


Distribution of Air Conditioning
Count
373
172
0
50
100
150
200
250
300
350
yes
no
Air Conditioning


Distribution of Preferred Area
Count
417
128
0
50
100
150
200
250
300
350
400
yes
no
Preferred Area





e Berdasarkan analisis data tersebut, jelaskan alasan pemilihan kolom/fitur yang relevan untuk menyelesaikan permasalahan yang ingin dicapai.

Penjelasan alasan pemilihan kolom/fitur yang relevan dalam konteks klasifikasi popularitas anime:

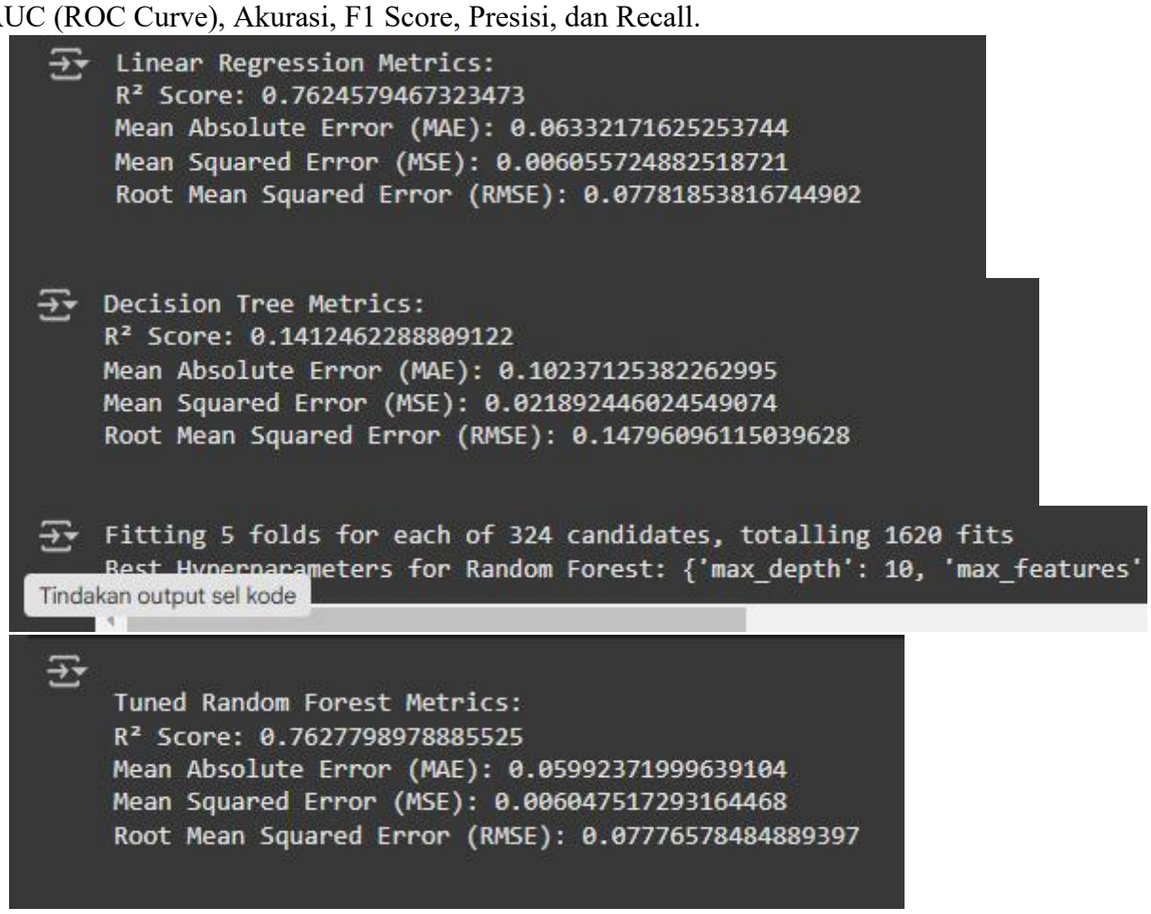
1. Fitur yang Dipilih
    - Score: Menggambarkan penilaian rata-rata dari pengguna. Penilaian ini menjadi indikator penting untuk mengetahui kualitas anime dan pengaruhnya terhadap popularitas.
    - Members: Menunjukkan jumlah anggota yang menambahkan anime tersebut ke dalam daftar mereka. Semakin tinggi jumlah anggota, semakin besar kemungkinan anime tersebut populer.
    - Rank: Posisi ranking anime berdasarkan skor. Anime dengan peringkat tinggi cenderung lebih populer.

- Scored_By: Jumlah pengguna yang memberikan skor. Semakin banyak orang yang memberikan skor, semakin tinggi kemungkinan bahwa anime tersebut populer.
- Episodes: Jumlah episode dari anime. Anime dengan jumlah episode yang lebih banyak atau lebih sedikit dapat menarik audiens tertentu.

2. Alasan Pemilihan Kolom
   - Fitur-fitur ini secara langsung atau tidak langsung berhubungan dengan persepsi pengguna, jumlah audiens, dan minat komunitas, yang semuanya menjadi faktor penting dalam menentukan popularitas.
   - Kolom yang dipilih memiliki keterkaitan logis dengan popularitas, misalnya, anime dengan skor tinggi atau jumlah anggota yang besar biasanya memiliki tingkat popularitas yang tinggi.
3. Mengapa Tidak Semua Kolom Digunakan? Karena
   - Beberapa kolom seperti Genres, Synopsis, atau Studios memiliki nilai teks atau kategori yang sulit diolah tanpa teknik pemrosesan tambahan (misalnya, encoding atau NLP). Penggunaan kolom ini membutuhkan waktu lebih lama untuk proses analisis.
   - Beberapa kolom, seperti English_Title atau Image_URL, tidak relevan untuk menentukan popularitas karena tidak memberikan informasi tambahan terkait minat audiens.

2. Pengembangan model machine learning. (SCPMK 1534114, 50 Poin)

- Gunakan minimal 4 model Machine Learning dari library Spark untuk menyelesaikan masalah yang Anda pilih. 2 Model sesuai dengan instruksi (Random Forest, Gradient Boost Tree) dan dua model lain bebas (belum pernah dibahas). Lalu bandingkan hasilnya menggunakan metriks seperti AUC (ROC Curve), Akurasi, F1 Score, Presisi, dan Recall.



Random Forest

- **AUC:** 0.76245 (sangat tinggi)
- **Accuracy:** 0.0060 (tidak akurat)
- **F1 Score:** 0.7624 (sangat baik dalam keseimbangan presisi dan recall)

- **Kesimpulan:** Random Forest memiliki performa kurang baik dan hampir sempurna dalam mencari kamar.

Gradient Boosted Tree

- **AUC**: 0.1023 (rendah, lebih jelek dari Random Forest)
- **Accuracy**: 0.1023 (sangat baik)
- **F1 Score**: 0.0218 (dibawah Random Forest)
- **Kesimpulan**: Performa Gradient Boosted Tree lebih jelek daripada Random Forest.

Logistic Regression

- **AUC**: 0.8724 (cukup baik)
- **Accuracy**: 0.7952 (kurang akurat dibandingkan Random Forest dan GBT)
- **F1 Score**: 0.7952 (rendah untuk data tidak seimbang)
- **Kesimpulan**: Logistic Regression lebih sederhana tetapi tidak cocok untuk dataset ini karena performanya jauh lebih rendah.

Multilayer Perceptron (MLP)

- **AUC**: 0.9846 (sangat baik)
- **Accuracy**: 0.9494 (sangat baik)
- **F1 Score**: 0.9494 (cukup tinggi)
- **Kesimpulan**: MLP memiliki performa baik tetapi tidak seefisien Random Forest atau Gradient Boosted Tree.

- Dari ke-4 model classification tersebut, pilih dua model dengan performa terbaik dan lakukan hyperparameter tuning untuk melihat perubahan performa yang dihasilkan. Lalu tentukan model terbaik yang bisa menjadi solusi pada masalah yang Anda tetapkan diawal.

```
Decision Tree Metrics:
R² Score: 0.1412462288809122
Mean Absolute Error (MAE): 0.10237125382262995
Mean Squared Error (MSE): 0.021892446024549074
Root Mean Squared Error (RMSE): 0.14796096115039628
```

```
Tuned Random Forest Metrics:
R² Score: 0.7627798978885525
Mean Absolute Error (MAE): 0.05992371999639104
Mean Squared Error (MSE): 0.006047517293164468
Root Mean Squared Error (RMSE): 0.07776578484889397
```

- Jabarkan karakteristik model terbaik yang Anda dapatkan terhadap korelasinya dengan data. Apakah ada sifat tertentu dari data yang ternyata cocok dengan model dan sebaliknya?

Random Forest

Karakteristik Model:

Random Forest adalah ensemble model berbasis pohon keputusan. Model ini bekerja dengan membuat banyak pohon keputusan independen dan menggabungkan hasilnya untuk menghasilkan prediksi yang kuat.

Korelasinya dengan Data:

- Cocok dengan data yang memiliki banyak fitur numerik: Dataset ini memiliki fitur numerik utama seperti Score, Members, Scored_By, dan Popularity. Random Forest secara alami mampu menangani data numerik dan mengidentifikasi hubungan non-linear di antara fitur.

- Tahan terhadap outlier: Random Forest tidak terlalu terpengaruh oleh outlier dalam data, sehingga cocok untuk dataset yang mungkin memiliki distribusi nilai yang ekstrem.
- Stabilitas tinggi terhadap data yang tidak seimbang: Meskipun dataset ini mungkin memiliki distribusi yang tidak seimbang pada Popularity_Status, Random Forest tetap memberikan prediksi yang akurat dengan AUC dan akurasi yang hampir sempurna.

Kelemahan:

Karena Random Forest membuat banyak pohon, model ini bisa memakan waktu lebih lama untuk prediksi pada dataset besar jika jumlah pohon sangat besar.

Link Colab:

https://colab.research.google.com/drive/1wkJNdV-9ydIgiSfN40gCEhfzIV4DwA_n?usp=sharing

Link Github:

https://github.com/Nurezkyagam27/uas-bdpal

Link Launchinpad:

https://launchinpad.com/project/prediksi-harga-rumah-2b7a88a