PREDICTION OF HOUSE PRICES IN CALIFORNIA USING LINEAR REGRESSION AND DECISION TREE REGRESSION

Yohanes Wiliam Hadiprojo

Information System Study Program
Multimedia Nusantara University
Tangerang, Indonesia
yohanes.wiliam@student.umn.ac.d

***Abstract*— Prediksi harga rumah sangat dibutuhkan bagi stakeholder real estate, seperti pembeli rumah, investor,, dan pembuat kebijakan. Studi ini menyediakan analisis dari beberapa pemodelan prediktif, yaitu Linear Regression dan Decision Tree Regression. Dengan melakukan dataset yang diperoleh dari Biro Sensus AS, kami melakukan processing, seperti imputasi missing value, transformasi fitur, normalisasi , dan encoding dalam mempersiapkan data untuk dianalisis. Pemodelan tersebut akan dinilai menurut nilai MSE dan R-Squared yang menunjukan hasil bahwa model Decision Tree Regression memperlihatkan sedikit peningkatan jika dibandingkan Linear Regression yang akan mengidentifikasikan hubungan kompleks dan tidak linear pada data. Visualisasi pada Decision Tree Regression memberikan wawasan tambahan. Prediksi ini memperlihatkan Linear Regression memberikan kesederhanaan dan interpretasi sementara Decision Tree memberikan akurasi lebih bagus. Studi ini memperlihatkan mengenai potensi machine learning dalam menciptakan strategi informasi dan menginformasikan penyempurnaan kebijakan dengan tujuan perkembangan dan stabilitas ekonomi.**

I. INTRODUCTION

Prediksi harga rumah memiliki tantangan besar yang memiliki faktor luas pada ekonomi, perencanaan kota, dan keuangan. Hal tersebut sangat melekat dengan real estate seperti pada perumahan di California yang sangat memerlukan penganalisaan akurat. Harga perumahan tersebut sangat dikaitkan dengan kondisi perekonomian serta demografi. Studi ini akan memberikan wawasan dengan beberapa model prediktif dalam memberikan prediksi harga yang akurat.

Penelitian ini ditujukan sebagai alat pembantu dalam pengambilan keputusan bagi stakeholder dalam pengambilan keputusan. Pembeli dapat menggunakannya sebagai penginformasi keputusan pembeliannya, Investor dapat menggunakannya untuk membantu strategi portofolio, Untuk pembuat kebijakan dapat digunakan sebagai pendukung kebijakan pada perumahan di California maupun pembangunan perekonomian kedepannya. Studi ini akan membandingkan kedua model yang sangat efektif dalam pemodelan harga rumah California.

Pada pemodelan Linear Regression terdapat kemungkinan pada hubungan linear pada prediktor dan variabel target tidak dapat memberikan penangkapan kompleksitas interaksi dan ketidaklinearannya. Jika dibandingkan dengan Decision Tree Regression memberikan wawasan yang lebih luas pada struktur data. Studi ini akan memberikan evaluasi model yang berguna.

Studi ini memberikan penguraian dataset serta preprocessing data dan proses pemodelan algoritma Linear Regression dan Decision Tree Regression serta pemberian hasil prediksi tersebut. Penelitian ini akan memberikan keterlibatan besar pemahaman mengenai pemodelan pada real estate dan penginformasian mengenai harga pasar pada perumahan California.

II. Literature Review

Data yang digunakan merupakan data set open source pada

website kaggle. Dataset yang digunakan bernama California

House Price. Dataset tersebut memberikan informasi mengenai:

1. koordinat geografis: Longitude dan Latitude
2. Atribut perumahan: Housing Median Age, Total Rooms, danTotal Bedrooms.
3. Populasi metriks: Population & Households
4. Indikator ekonomi: Meidan income & Median House Value
5. Location Specifics: Ocean Proximity

Informasi dataset tersebut memberikan informasi yang berkontribusi untuk prediksi ini.

Upaya untuk memprediksi harga real estat memiliki sejarah interdisipliner yang kaya, mencakup ilmu ekonomi, ilmu data, dan studi perkotaan. Penelitian awal oleh Quigley (1995) menekankan pengaruh faktor ekonomi terhadap pasar perumahan, dengan mencatat pentingnya pendapatan, tingkat pekerjaan, dan suku bunga dalam membentuk dinamika harga. Makalah penting oleh Case dan Shiller (1989) memperkenalkan konsep bahwa pasar perumahan memiliki komponen prediksi yang kuat berdasarkan harga historis, sebuah gagasan yang telah didukung dan dibantah dalam penelitian selanjutnya.

Kemunculan pembelajaran mesin telah mengantarkan era baru pemodelan prediktif. Regresi linier telah digunakan secara luas dalam prediksi harga rumah karena mudah diinterpretasikan dan sederhana (Zillow Research, 2018). Namun, kinerja prediktifnya sering kali dikalahkan oleh model yang lebih kompleks yang dapat menangkap hubungan non-linear (Park dan Bae, 2015).

Studi terbaru semakin banyak beralih ke Decision Tree Regression dan varian ensembelnya, seperti Random Forest dan Gradient Boosting, untuk memprediksi harga rumah. Model-model ini telah terbukti mengungguli model linier dalam menangkap sifat pasar real estat yang memiliki banyak aspek (Li et al., 2020). Karya Geurts dkk. (2006) tentang algoritme Random Forest menyoroti kekuatannya dalam menangani set data berdimensi tinggi yang biasa ditemui dalam studi perumahan.

Model hibrida juga telah mendapatkan perhatian. Sebagai contoh, Petkov (2017) menggabungkan beberapa teknik pembelajaran mesin untuk meningkatkan akurasi prediksi, yang mengindikasikan potensi dari pendekatan ansambel. Integrasi analisis spasial dengan model prediktif, seperti yang dieksplorasi oleh Helbich dkk. (2013), semakin menunjukkan kompleksitas faktor penentu harga perumahan, seperti fitur geografis dan kedekatan.

Analisis komparatif dari model-model ini dalam konteks California masih belum banyak didokumentasikan. California merupakan pasar yang unik karena ukuran, keragaman ekonomi, dan faktor lingkungannya, sehingga menjadikannya studi kasus yang ideal untuk teknik pemodelan prediktif tingkat lanjut (Saks dan Wachter, 2018).

Penelitian ini bertujuan untuk berkontribusi pada literatur yang ada dengan tidak hanya membandingkan pendekatan tradisional dan pembelajaran mesin untuk memprediksi harga rumah di California, tetapi juga dengan memeriksa implikasi dari model prediktif ini dalam konteks pengambilan keputusan ekonomi.

III. METHODOLOGY

Pada bagian ini akan menjelaskan penguraian langkah pemprosesan data dan mempersiapkannya untuk model linear regression dalam mengoptimalkan akurasi dalam prediksi harga rumah.

Tahapan pada proses pre processing termasuk:

a. DATA CLEANSING

Langkah pertama ini untuk menghilangkan Missing values yang dapat menyebabkan hasil bias. Tujuan dari cleansing data ini untuk identifikasi entri yang hilang dari pemeriksaan isnull(). Pada fitur numerik tersebut missing values diperhitungkan dengan nilai median dan untuk fitur kategorikal modus dipakai. Ini akan membenarkan inputasi tidak signifikan yang menggeser distribusi fitur.

b. HANDLING OUTLIERS

Outlier dapat memberikan dampak pada hasil model yang tidak sebanding. Langkah ini akan menggunakan metode IQR dalam identifikasi outlier. Outlier diadaptasikan dengan threshold value jika menunjukan sebagai kesalahan yang akan dihapus dalam mencegah kemiringan.

c. FORMATING

Langkah ini memper standarkan format teks untuk variabel kategorikal, memastikan nilai numerik mempunyai tempat desimal yang konsisten dan mengkonversinya dalam bentuk datetime.

d. NORMALIZATION

penskalaan Min-Max digunakan untuk menskalakan fitur numerik pada [0,1] dimana hal tersebut akan sangat penting dalam kedua pemodelan tersebut

e. BINNING

langkah tersebut dibutuhkan dalam mengubah variabel kontinu menjadi kategorik.

f. ENCODING

Variabel kategorikal dikodekan dalam memungkinkan dalam melakukan melakukan kedua pemodelan. One hot-encoding dilakukan pada variabel ocean_proximity, yang mengubah menjadi serangkain kolom biner yang menggantikan keberadaan kategori.

g. GROUPING

Langkah ini diperlukan untuk melakukan analisis pada tingkat yang lebih rinci dengan contoh pada variabel ocean_proximity yang diperlukan dalam analisa tren pada coastal & inland area yang akan memberikan pernyataan subgroup pattern pada data untuk memperoleh wawasan yang lebih dalam.

h. DATA SPLITTING

Data spitting merupakan langkah akhir yang akan membagi dataset menjadi training & set test pada rasio 80/20 yang melakukan pengambilan sampel bertingkat.

IV. RESULT & EXPLANATION

A. DATA PREPROCESSING & EDA (EXPLORATORY DATA ANALYSIS)

menampilkan 10 baris pertama dari set data berjudul "California House Price.csv". Dataset ini mencakup beberapa kolom dengan berbagai jenis data yang berkaitan dengan perumahan di California. Berikut adalah penjelasan dari kolom-kolom tersebut:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
0	-122.23	37.88	41	880	129.0	322	126	8.3252
1	-122.22	37.86	21	7099	1106.0	2401	1138	8.3014
2	-122.24	37.85	52	1467	190.0	496	177	7.2574
3	-122.25	37.85	52	1274	235.0	558	219	5.6431
4	-122.25	37.85	52	1627	280.0	565	259	3.8462
5	-122.25	37.85	52	919	213.0	413	193	4.0368
6	-122.25	37.84	52	2535	489.0	1094	514	3.6591
7	-122.25	37.84	52	3104	687.0	1157	647	3.1200
8	-122.26	37.84	42	2555	665.0	1206	595	2.0804
9	-122.25	37.84	52	3549	707.0	1551	714	3.6912

ocean_proximity	median_house_value
NEAR BAY	452600
NEAR BAY	358500
NEAR BAY	352100
NEAR BAY	341300
NEAR BAY	342200
NEAR BAY	269700
NEAR BAY	299200
NEAR BAY	241400
NEAR BAY	226700
NEAR BAY	261100

1. longitude: Koordinat bujur lokasi rumah..
2. latitude: Koordinat lintang lokasi rumah.
3. housing_median_age: Usia rata-rata rumah di blok tersebut; ini dapat mencerminkan era pembangunan atau mengindikasikan apakah area tersebut memiliki rumah yang lebih baru atau lebih tua.
4. total_rooms: Jumlah total ruangan di blok rumah.
5. total_bedrooms: Jumlah kamar tidur di blok rumah.
6. population: households: Jumlah total orang yang tinggal di blok tersebut.
7. rumah tangga: Jumlah total rumah tangga atau unit-unit tempat tinggal yang berbeda di dalam blok tersebut.
8. median_income: Pendapatan rata-rata rumah tangga di dalam blok, biasanya diukur dalam puluhan ribu. Sebagai contoh, nilai 8.3252 akan diterjemahkan menjadi $83.252.
9. ocean_proximity:Variabel kategorikal yang menunjukkan kedekatan rumah dengan laut, dengan kategori seperti 'DEKAT TELUK' dalam dataset ini.
10. median_house_value: Nilai rata-rata rumah di blok, yang biasanya merupakan variabel target untuk prediksi dalam model harga rumah.

```
In [35]: house.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  int64
 3   total_rooms         20640 non-null  int64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  int64
 6   households          20640 non-null  int64
 7   median_income       20640 non-null  float64
 8   ocean_proximity     20640 non-null  object
 9   median_house_value  20640 non-null  int64
dtypes: float64(4), int64(5), object(1)
memory usage: 1.6+ MB
```

Ini menunjukkan bahwa DataFrame house adalah objek pandas dengan 20.640 entri, diindeks dari 0 hingga 20.639.

Ada total 10 kolom dalam DataFrame. Output menampilkan nama setiap kolom, jumlah nilai non-null (tidak hilang), dan tipe data (Dtype) untuk setiap kolom:

1. longitude: 20.640 entri non-null, tipe data float64.
2. latitude: 20.640 entri non-null, tipe data float64.
3. housing_median_age: 20.640 entri non-null, tipe data int64.
4. total_rooms: 20.640 entri non-null, tipe

data int64.

5. total_bedrooms: 20.433 entri non-null, tipe data float64 (menunjukkan ada beberapa nilai yang hilang).
6. population: 20.640 entri non-null, tipe data int64.
7. households: 20.640 entri non-null, tipe data int64.
8. median_income: 20.640 entri non-null, tipe data float64.
9. ocean_proximity: 20.640 entri non-null, tipe data object (biasanya string atau tipe campuran).
10. median_house_value: 20.640 entri non-null, tipe data int64.

DataFrame berisi empat kolom float64 (longitude, latitude, total_bedrooms, median_income), lima kolom int64 (housing_median_age, total_rooms, population, households, median_house_value), dan satu kolom object (ocean_proximity).

```
In [37]: # Handling missing values
         house['total_bedrooms'].fillna(house['total_bedrooms'].median(), inplace=Tru
```

Baris kode ini mengatasi nilai yang hilang pada kolom total_bedrooms dengan mengisinya menggunakan nilai median dari kolom tersebut. Operasi ini dilakukan secara langsung pada DataFrame asli (inplace=True).

```
In [38]: # Kolom yang akan ditransformasi logaritmik
         logtransform = ['total_rooms', 'total_bedrooms', 'population', 'households', 'median_in
```

Di sini, daftar kolom yang akan di-transformasi secara logaritmik ditentukan. Transformasi logaritmik sering digunakan untuk mengurangi kemiringan dalam data. Fungsi .apply() digunakan dengan fungsi lambda untuk menerapkan fungsi np.log() ke setiap kolom yang ditentukan.

```
In [38]: # Kolom yang akan ditransformasi logaritmik
         logtransform = ['total_rooms', 'total_bedrooms', 'population', 'households', 'median_income']

In [39]: # Menerapkan transformasi logaritmik pada kolom tertentu
         house[logtransform] = house[logtransform].apply(lambda x: np.log(x + 1))
```

Di sini, daftar kolom yang akan di-transformasi secara logaritmik ditentukan. Transformasi logaritmik sering digunakan untuk mengurangi kemiringan dalam data. Fungsi .apply() digunakan dengan fungsi lambda untuk menerapkan fungsi np.log() ke setiap kolom yang ditentukan

```
In [40]: # Normalization using Min-Max Scaling
         scaler = MinMaxScaler()
         numerical_cols = house.select_dtypes(include=['float64', 'int64']).columns
         house[numerical_cols] = scaler.fit_transform(house[numerical_cols])
```

Blok ini menginisialisasi Skala Min-Max, memilih kolom numerik, dan menerapkan skala ke masing-masing, mengubah nilai ke rentang antara 0 dan 1.

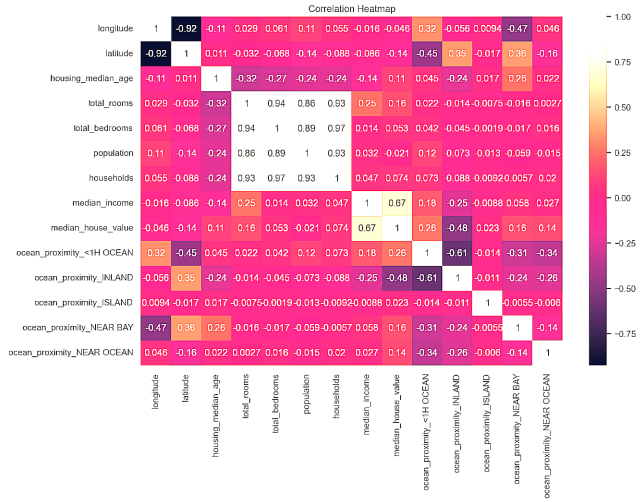
```
In [41]: # Encoding for categorical data using one-hot encoding
         house_encoded = pd.get_dummies(house, columns=['ocean_proximity'])
```

Fungsi pd.get_dummies() digunakan untuk melakukan one-hot encoding pada kolom ocean_proximity, membuat kolom biner baru untuk setiap kategori dalam kolom tersebut.

```
In [42]: # Cek apakah ada nilai NaN dalam data
         print("Cek nilai NaN sebelum pemodelan:")
         print(house_encoded.isna().sum())

Cek nilai NaN sebelum pemodelan:
longitude                   0
latitude                    0
housing_median_age          0
total_rooms                 0
total_bedrooms              0
population                  0
households                  0
median_income               0
median_house_value          0
ocean_proximity_<1H OCEAN   0
ocean_proximity_INLAND      0
ocean_proximity_ISLAND      0
ocean_proximity_NEAR BAY    0
ocean_proximity_NEAR OCEAN  0
dtype: int64
```

Kode ini memeriksa nilai NaN (Not a Number) pada DataFrame house_encoded, yaitu DataFrame setelah encoding. Rantai metode .isna().sum() digunakan untuk menghitung jumlah NaN pada setiap kolom. Output menunjukkan angka nol pada seluruh kolom, menandakan tidak ada lagi nilai yang hilang dalam data.



kemungkinan besar dataset harga rumah di California. Matriks korelasi adalah tabel yang menunjukkan koefisien korelasi antara variabel. Setiap sel dalam tabel menunjukkan korelasi antara dua variabel. Nilainya berkisar antara -1 hingga 1. Jika dua variabel memiliki korelasi yang tinggi (mendekati 1 atau -1), ini berarti ketika satu variabel berubah, variabel lain juga cenderung berubah dalam arah tertentu. Jika korelasi mendekati 0, ini berarti tidak ada hubungan linier antara variabel tersebut.

1. Warna Merah: Menunjukkan korelasi positif, yang berarti bahwa saat satu variabel meningkat, variabel lain juga cenderung meningkat.
2. Warna Biru: Menunjukkan korelasi negatif, yang berarti bahwa saat satu variabel meningkat, variabel lain cenderung menurun.
3. Intensitas Warna: Intensitas warna sesuai dengan kekuatan korelasi. Warna yang lebih gelap (baik merah atau biru) menunjukkan korelasi yang lebih kuat, sementara warna yang lebih terang menunjukkan korelasi yang lebih lemah.

beberapa yang dapat diamati dari heatmap tersebut:

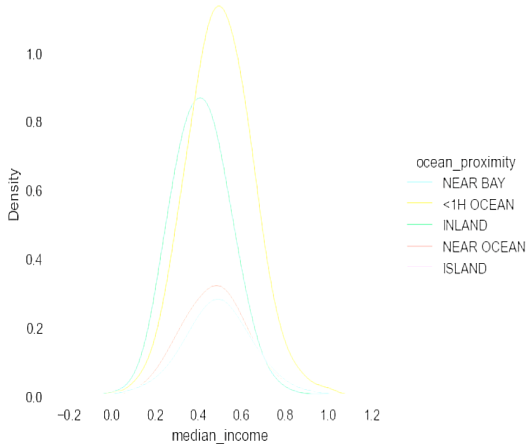
1. latitude dan longitude: Ada korelasi negatif yang kuat antara latitude dan longitude, mungkin menunjukkan pola atau tren geografis di dalam California.

2. housing_median_age: Tampaknya ada sedikit atau tidak ada korelasi antara housing_median_age dengan longitude dan latitude, menunjukkan bahwa usia median perumahan tidak tergantung pada lokasi di dalam California.

3. total_rooms dan total_bedrooms: Ada korelasi positif yang sangat tinggi antara total_rooms dan total_bedrooms, menunjukkan bahwa rumah dengan lebih banyak ruangan juga cenderung memiliki lebih banyak kamar tidur.

4. median_income dan median_house_value: Median_income memiliki korelasi positif yang kuat dengan median_house_value, yang bisa menyiratkan bahwa area dengan median_income yang lebih tinggi juga memiliki harga rumah yang lebih mahal.

5. ocean_proximity: Variabel kategorikal ocean_proximity telah diubah menjadi beberapa kolom biner, masing-masing mewakili sebuah kategori. Korelasi antara kolom biner baru ini dengan variabel lain dapat menunjukkan bagaimana kedekatan dengan lokasi samudra yang berbeda mungkin berhubungan dengan karakteristik dan harga perumahan.



Plot kepadatan ini menggambarkan distribusi median_income berdasarkan kedekatan dengan laut (ocean_proximity). Setiap garis pada plot tersebut mewakili sebuah kategori kedekatan dengan laut, seperti NEAR BAY, <1H OCEAN, INLAND, NEAR OCEAN, dan ISLAND. Kepadatan pada sumbu y menunjukkan seberapa sering nilai median_income muncul dalam dataset, sementara sumbu x menunjukkan rentang nilai median_income yang telah dinormalisasi.

NEAR BAY: Garis ini menunjukkan distribusi median_income untuk daerah yang dekat dengan teluk.

<1H OCEAN: Menunjukkan distribusi untuk daerah yang kurang dari satu jam dari laut.

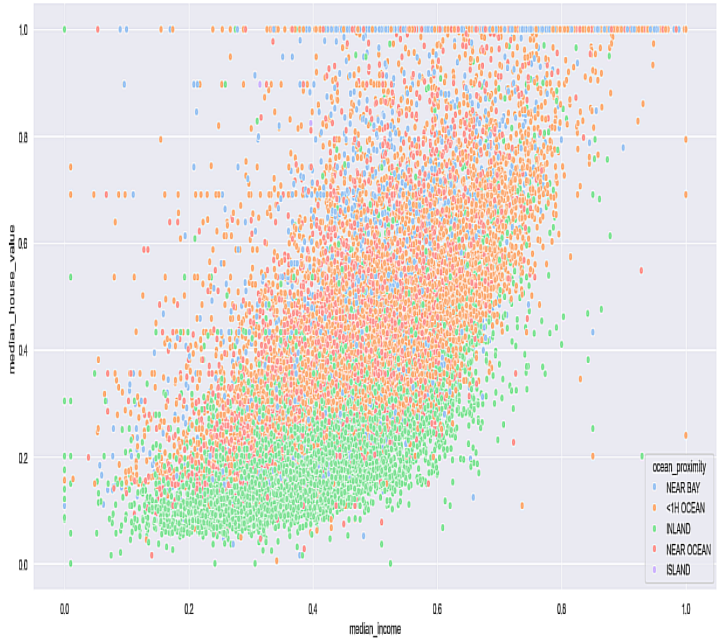
INLAND: Menunjukkan distribusi untuk daerah yang berada di daratan atau jauh dari laut.

NEAR OCEAN: Menunjukkan distribusi untuk daerah yang dekat dengan pantai laut.

ISLAND: Menunjukkan distribusi untuk daerah yang merupakan pulau.

Parameter bw_adjust=2 digunakan untuk membuat garis kepadatan lebih halus, yang bisa membantu dalam menunjukkan tren umum daripada variasi kecil dalam data. Palette "pastel" memberikan warna lembut yang berbeda untuk setiap kategori ocean_proximity, yang membantu dalam membedakan antara kelompok-kelompok tersebut secara visual.

Dari plot ini, kita dapat menganalisis bagaimana median pendapatan berbeda-beda tergantung pada kedekatan dengan laut, yang bisa memberikan wawasan tentang faktor-faktor ekonomi yang mungkin mempengaruhi nilai pasar perumahan di berbagai lokasi di California.



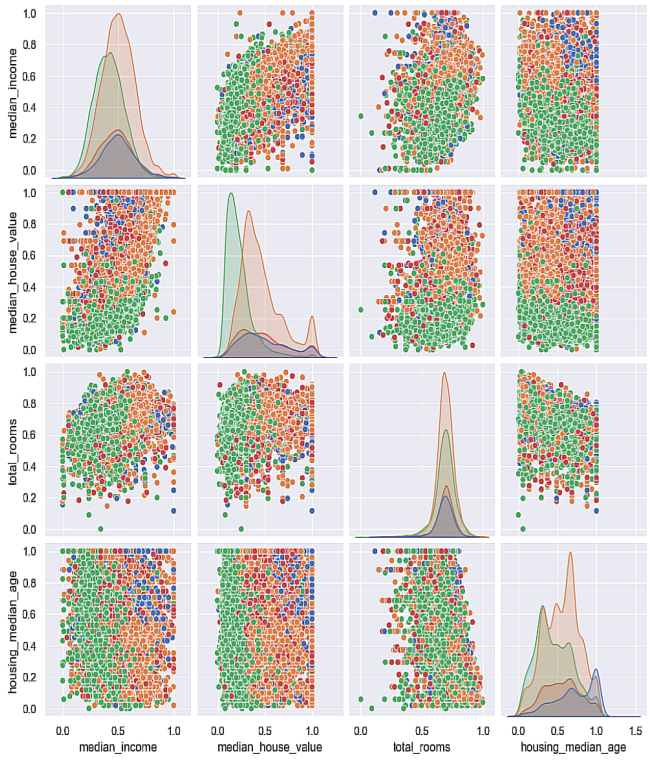
Plot sebar ini menampilkan hubungan antara pendapatan median dari rumah tangga dengan nilai median rumah. Setiap titik pada plot mewakili sebuah blok rumah. Sumbu x menunjukkan median_income, dan sumbu y menunjukkan median_house_value. Variabel ocean_proximity digunakan untuk memberi warna pada titik-titik tersebut, dengan warna yang berbeda menunjukkan kategori yang berbeda seperti NEAR BAY (dekat teluk), <1H OCEAN (kurang dari satu jam dari laut), INLAND (daratan), NEAR OCEAN (dekat laut), dan ISLAND (pulau).

1. Ada hubungan positif antara median_income dan median_house_value, di mana wilayah dengan pendapatan median yang lebih tinggi cenderung memiliki nilai rumah yang lebih tinggi

pula.

2. Kedekatan dengan laut tampaknya berpengaruh pada nilai rumah; contohnya, rumah-rumah yang terletak di NEAR BAY dan NEAR OCEAN mungkin memiliki nilai yang lebih tinggi dibandingkan dengan yang di INLAND.
3. Titik-titik data terdistribusi luas, menunjukkan variasi nilai rumah yang signifikan di berbagai kisaran pendapatan dan lokasi relatif terhadap laut.

Palet "pastel" digunakan untuk memberikan tampilan visual yang menarik dan memudahkan identifikasi grup berdasarkan kedekatan dengan laut. Plot seperti ini sangat berguna untuk menganalisis tren dan pola dalam data yang besar dan dapat membantu dalam pemilihan fitur untuk model prediktif harga rumah.



Pairplot ini memvisualisasikan hubungan antara kombinasi berbagai variabel. Setiap baris dan kolom dari grafik mengacu pada satu variabel. Di mana baris dan kolom ini bertemu, hubungan antara dua variabel tersebut ditampilkan:

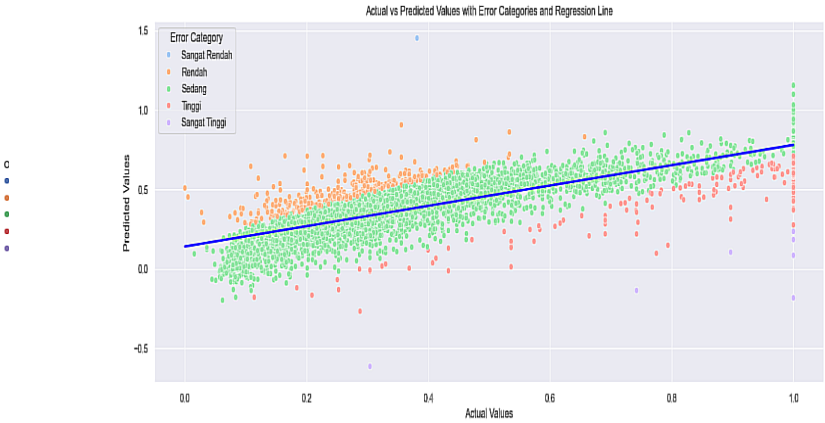
Pada diagonal utama, distribusi dari setiap variabel ditampilkan menggunakan histogram atau grafik kepadatan.

Pada sel off-diagonal, scatter plots menunjukkan hubungan antara pasangan variabel. Sebagai contoh, plot pada baris pertama dan kolom kedua menunjukkan hubungan antara median_income dan median_house_value.

Variabel ocean_proximity mempengaruhi warna dari titik data di scatter plots, yang memungkinkan kita untuk melihat bagaimana hubungan antara variabel numerik mungkin berbeda berdasarkan kedekatan dengan laut. Dari plot ini, kita dapat mengamati, misalnya, apakah rumah yang lebih dekat dengan laut cenderung memiliki nilai atau pendapatan median yang lebih tinggi, atau bagaimana total ruangan dan usia rumah mungkin berhubungan dengan kedekatan dengan laut dan nilai rumah.

B. MACHINE LEARNING RESULT

1. linear regression



Plot ini menggambarkan seberapa baik model prediksi bekerja dengan membandingkan nilai yang diprediksi oleh model dengan nilai sebenarnya dari data. Titik-titik data diwarnai berbeda berdasarkan kategori kesalahan prediksi, seperti 'Sangat Rendah', 'Rendah', 'Sedang', 'Tinggi', dan 'Sangat Tinggi'. Warna pastel digunakan untuk membedakan antara kategori kesalahan ini.

Garis regresi biru menunjukkan tren prediksi model. Idealnya, titik-titik data harus berada dekat

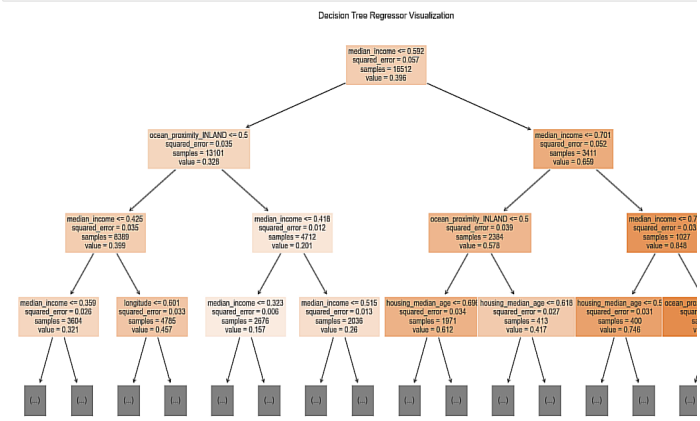
dengan garis ini, yang akan menunjukkan bahwa nilai yang diprediksi dekat dengan nilai aktual. Jika titik-titik tersebar jauh dari garis regresi, ini menunjukkan adanya kesalahan dalam prediksi.

```
In [61]: print(f"Decision Tree Regressor Mean Squared Error: {mse_tree}")
         print(f"Decision Tree Regressor R-squared: {r2_tree}")

Decision Tree Regressor Mean Squared Error: 0.02040255137839828
Decision Tree Regressor R-squared: 0.633760924359849
```

DecisionTreeRegressor. MSE sebesar 0.02040255137839828 menunjukkan bahwa, rata-rata, kuadrat kesalahan prediksi adalah 0.0204. Nilai R-squared sebesar 0.633760924359849 menunjukkan bahwa sekitar 63.38% variabilitas dalam median_house_value dapat dijelaskan oleh model pohon keputusan yang telah dilatih.

Nilai MSE yang rendah dan R-squared yang tinggi menunjukkan bahwa model memiliki performa yang baik dalam sampel data uji yang diberikan.



Decision Tree Regressor Mean Squared Error: 0.02040255137839828
Decision Tree Regressor R-squared: 0.633760924359849

Model Decision Regression Treemenghasilkan nilai R-squared yang lebih tinggi, yang mengindikasikan kecocokan yang lebih baik terhadap data dan kemampuan yang lebih baik untuk menjelaskan varians harga rumah.

Pada model Regresi Linier, meskipun lebih sederhana, menunjukkan keterbatasan dalam menangkap kompleksitas data. Model Regresi Pohon Keputusan menghasilkan nilai R-squared yang lebih tinggi, yang mengindikasikan kecocokan yang lebih baik terhadap data dan kemampuan yang lebih baik untuk menjelaskan varians harga rumah.

Evaluasi Model:

Nilai MSE dan R-squared untuk kedua model dihitung, dengan model Decision Tree menunjukkan MSE yang lebih rendah dan R-squared yang lebih tinggi, yang menandakan prediksi yang lebih akurat dan kecocokan yang lebih baik secara keseluruhan.

Analisis Komparatif dan Peningkatan

pada Perbandingan Kinerja menunjukan performa superior model Decision Tree menunjukkan keefektifannya dalam menangkap hubungan yang kompleks dan tidak linier dalam data pasar perumahan. Selain itu pada Perbaikan model di masa depan dapat mencakup eksplorasi metode ensemble, rekayasa fitur lebih lanjut, dan kemungkinan memasukkan sumber data tambahan untuk meningkatkan daya prediksi model dan Temuan ini menawarkan wawasan yang berharga bagi para pemangku kepentingan di bidang real estat, menyediakan alat untuk pengambilan keputusan yang lebih tepat di pasar perumahan.