



# California Housing Price Prediction

MUHAMMAD AL-FARISY  
JCDS 1704 - PURWADHIKA DIGITAL TECHNOLOGY SCHOOL

# DAFTAR ISI :



# Business Problem



01



## Context

- California memiliki peluang besar untuk bisnis property.
- Agent properti menentukan harga rumah dengan cara membandingkan properti yang ingin dijual dengan properti lain disekitar daerah tersebut.
- Kenaikan harga rumah mencapai 25% dalam 5 tahun terakhir (1985-1989) dan mengalami penurunan harga secara drastic di tahun 1990 mencapai 25% penurunan harga rumah.
- Berdampak juga pada keuangan agent properti

02



## Problem Statement

Pengaruh dari kurang tepatnya dalam penentuan harga rumah ini berdampak pada penurunan penjualan agent properti yang menyebabkan harga tidak stabil dan customer tidak mampu membeli karena harganya yang naik secara drastic. Hal menjadi tantangan untuk agent properti untuk menentukan harga rumah yang dijual secara tepat.



## Business Problem



03



### Goals

- Agen properti memerlukan perangkat yang dapat membantu memprediksi harga rumah secara tepat.
- Dapat agen properti gunakan ketika client (pihak pemilik rumah) ingin menjual rumahnya dan agen properti sebagai perantara
- Memprediksi harga rumah yang sesuai berdasarkan faktor-faktor lokasi rumah, umur rumah, jumlah ruangan, jumlah kamar tidur, keramaian lingkungan rumah, kawasan perumahan elit atau tidak dan jarak rumah dengan lokasi wisata dengan menggunakan machine learning

## Business Problem



04



### Analytic Approach

1. Menganalisis data untuk dapat menemukan pola dari fitur-fitur yang ada, yang membedakan satu perumahan dengan yang lainnya.
2. Membangun model regresi yang dapat membantu agent properti untuk menyediakan sebuah perangkat prediksi harga rumah.
3. Menemukan fitur yang memiliki pengaruh besar terhadap target median house values agar model yang dibangun lebih sederhana dan memiliki tingkat akurat yang baik

05



### Metric Evaluation

**MAE ->** Rata rata selisih harga aktual dengan harga prediksi (US Dollar)

**MAPE -** Persentase rata-rata error antar harga aktual dengan harga prediksi dibagi harga actual x 100%

# Data Understanding



## **longitude**

Ukuran seberapa jauh ke barat sebuah rumah; nilai yang lebih tinggi lebih jauh ke barat

## **latitude**

Ukuran seberapa jauh ke utara sebuah rumah; nilai yang lebih tinggi lebih jauh ke utara

## **housing median age**

Usia rata-rata sebuah rumah dalam satu blok (angka yang lebih rendah adalah bangunan yang lebih baru)

## **total rooms**

Jumlah total ruangan dalam rumah didalam satu blok

## **total bedrooms**

Jumlah total kamar tidur dalam satu blok

## **population**

Jumlah total orang yang tinggal dalam satu blok

## **households**

Jumlah total rumah tangga / sekelompok orang yang tinggal dalam satu unit rumah di satu blok

# Data Understanding



## median income

Pendapatan rata-rata untuk rumah tangga dalam satu blok rumah (diukur dalam puluhan ribu Dolar AS)

## ocean proximity

Lokasi rumah dengan laut/laut

## median house value

Nilai median rumah untuk rumah tangga dalam satu blok (diukur dalam Dolar AS)

# DATA PREPARATION

## Total Baris dan Kolom

Total baris ada 14.448 dan total kolom ada 10

## Data Duplikate

Data Duplikat : 0

## Handling Missing Values

Total Data 100%  
Missing Values 0.948 %  
Data Bersih 99.052 %



## Mengganti Tipe Data

- Total\_rooms
- Total\_bedrooms
- Population
- Households

## Outlier

- housing\_median\_age yang diatas 51 tahun di hapus (Akan Menjadi Project Limitasi)
- Batas maksimal harga rumah 462200 US Dollar

## Korelasi

- longitude dan latitude (Negatif & Lemah)
- house\_median\_age, total\_rooms, total\_bedrooms, population, households (Positif & Lemah)
- median\_income memiliki (Positif & Sedang)



# California House City

EASY TO CHANGE COLORS, PHOTOS.

MODEL  
MACHINE  
LEARNING



# BASE MODEL

	MAE	MAPE
Testing Linear Regression	42086.464186	0.262159
Testing KNN	35686.775164	0.209156
Testing Decision Tree Regressor	41444.380665	0.250883
Testing RandomForest	29916.280343	0.184210
Testing XGBoost	29072.745965	0.176594

**Hasil Based Model** : Saat dilakukan prediksi pada testing, performa XGBoost memiliki MAE dan MAPE yang paling kecil sehingga memiliki kedekatan pada akurasi jika nilainya semakin kecil atau mendekati 0. XGBoost dipilih sebagai model akhir dan untuk meningkatkan performanya, untuk meningkatkan perfoma model XGB maka akan dilakukan hyperparameter tuning pada model XGBoost.

# MODEL TUNING

## Definisi XGBoost :

XGBoost adalah pustaka pembelajaran mesin pohon keputusan (GBDT) yang dapat diskalakan dan terdistribusi. Ini memberikan peningkatan pohon paralel dan merupakan perpustakaan pembelajaran mesin terkemuka untuk masalah regresi, klasifikasi, dan peringkat.

## Hyperparameter Tuning :

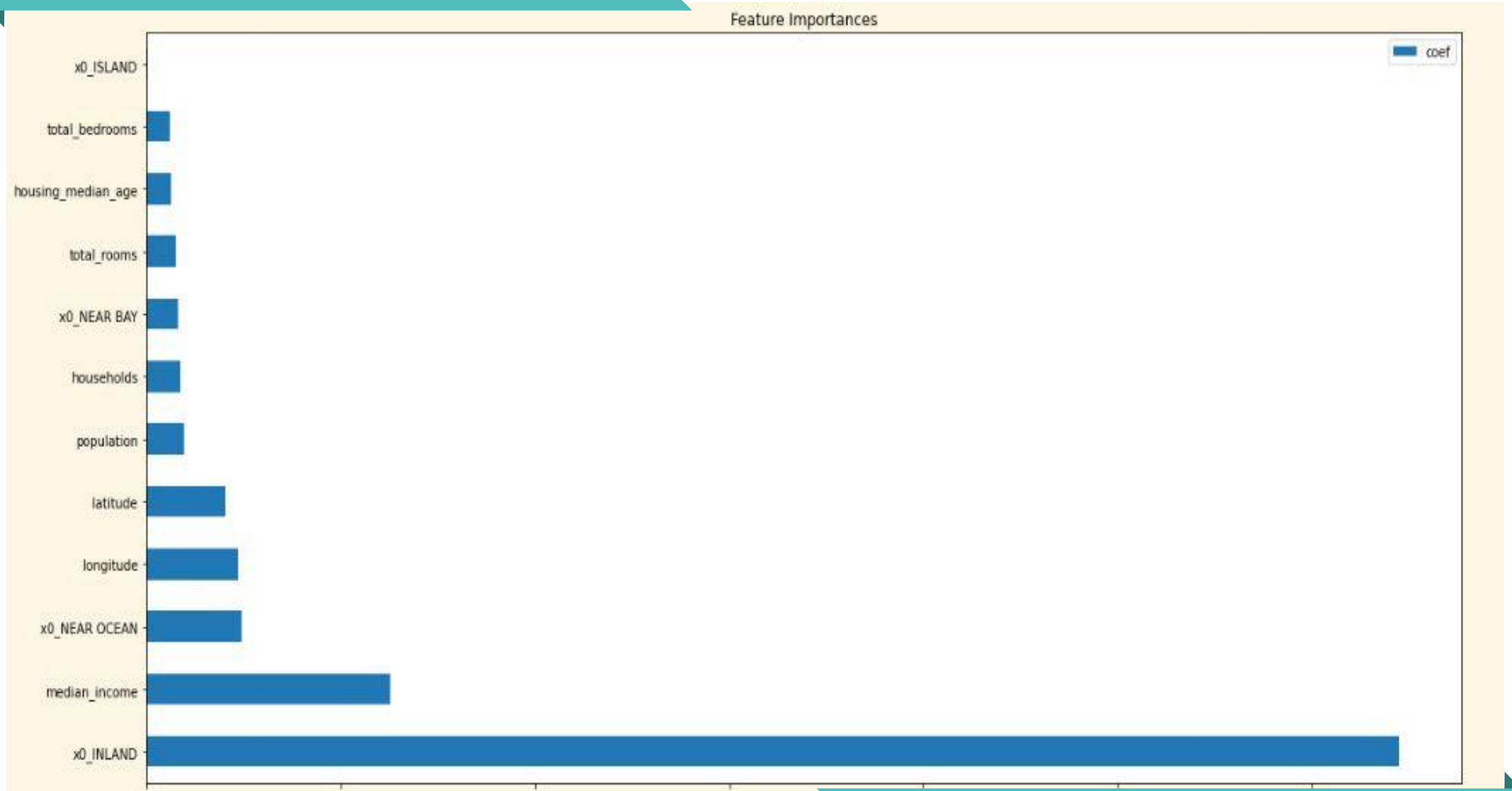
RandomizedSearchCV

	MAE	MAPE
XGB	28293.015355	0.168612

## Param\_distributions :

1. **max\_depth (Kedalaman pohon)** = `list(np.arange(1, 11))` -> max\_depth terbaik 10
2. **learning\_rate (kecepatan belajar)** = `list(np.arange(1, 100)/100)` -> learning rate terbaik 0.03
3. **n\_estimators (jumlah pohon dibangun)** = `list(np.arange(100, 201))` -> estimator terbaik 141
4. **Subsample (jumlah baris tiap pohon)** = `list(np.arange(2, 10)/10)` -> Subsample terbaik 6
5. **Gamma (node yang dipecah Ketika split)** = `list(np.arange(1, 11))` -> Gamma terbaik 10
6. **colsample\_bytree (jumlah feature tiap pohon)** = `list(np.arange(1, 10)/10)` -> colsample\_bytree terbaik 0.6
7. **reg\_alpha (regularization)** = `list(np.logspace(-3, 1, 10))` -> reg\_alpha terbaik 0.059948

# FEATURE IMPORTANCE



# PERBANDINGAN NON ML VS USE ML

	median_house_value	Harga_Prediksi_Rumah	Selisih_Harga
1	100000.0	104906.0	4906.0
2	285800.0	276521.0	9279.0

- Dari sample perbandingan data di atas, terdapat 2 index dengan nilai prediksi yang naik dan turun. Penjelasan sebagai berikut :

- Index 1 **Case Harga prediksi lebih tinggi** : Median\_house\_values bernilai 100000.0 US Dollar, Harga\_Prediksi\_Rumah bernilai 104906.0 US Dollar dan selisih antara harga aktual dengan harga prediksi 4906.0 US Dollar. Artinya lokasi perumahan yang dijual adalah perumahan di kawasan elit yang penduduknya memiliki median\_income yang tinggi dan berlokasi jauh dari laut sehingga harga dari prediksi mengalami kenaikan.

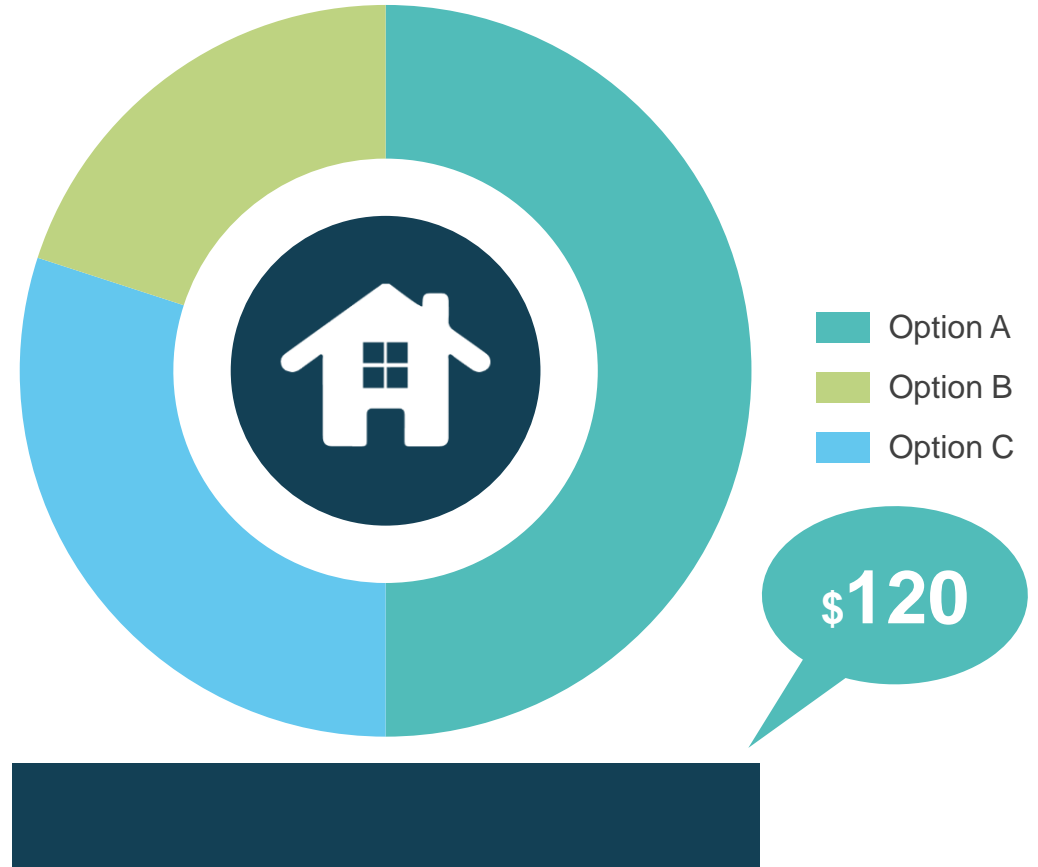
- Index 2 **Case Harga prediksi lebih rendah dari harga aktual** : Median\_house\_values bernilai 285800.0 US Dollar, Harga\_Prediksi\_Rumah bernilai 276521.0 US Dollar dan selisih antara harga aktual dengan harga prediksi 9279.0 US Dollar. Artinya lokasi perumahan yang dijual bukan dikawasan elit yang penduduknya tidak memiliki pendapatan yang tinggi dan berlokasi dekat dengan laut.



# Kesimpulan

Your Text Here

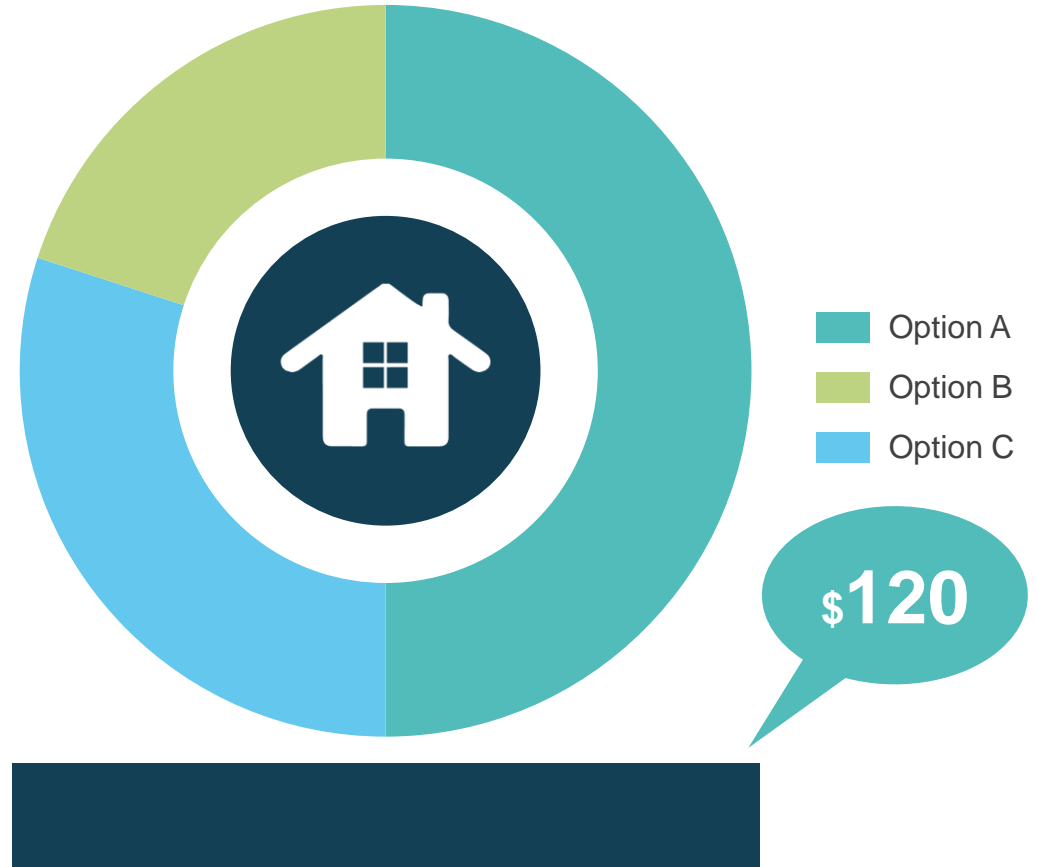
- Best model XGBoost
- Feature importance feature median\_income dan feature ocean\_proximity.
- MAE 28293.0 US Dollar
- MAPE 0.168612 (16 %)
- Project Limitation :
  - Age House Max 50
  - Max Price House 462200 US Dollar



# Saran

Your Text Here

- Penambahan feature yang berkaitan dengan fasilitas rumah (luas tanah dan bangunan, Parkiran mobil, Garasi dll)
- Mengembangkan metode tuning dengan GridSearch untuk mendapatkan hasil optimal
- Tidak disarankan menggunakan rumah sudah yang berdiri 50 tahun dan harga perumahan tidak lebih dari 462200 US Dollar.



THANK YOU

