Predictive Analytics in Real Estate: Leveraging Housing Data to Forecast Market Trends and Revitalization

Opportunities

By Ashish Mangal

<u>Abstract</u>

This study aims to explore the use of the "Housing & Ocean Proximity" dataset to predict key real estate market indicators. It focuses on forecasting housing prices, crime rates, and identifying areas for potential urban revamping. By integrating machine learning algorithms with comprehensive housing data, including neighborhood demographics and economic conditions, the study aims to offer actionable insights for investors, urban planners, and policy makers, contributing to a more responsive and informed approach in the dynamic real estate sector.

<u>Keywords</u>

Real Estate Analytics, Machine Learning, Housing Market trends, Urban Planning

Introduction

The rapid evolution of the real estate market, influenced by changing demographics, economic factors, and urbanization, necessitates advanced analytical approaches. This paper delves into the potential of machine learning in transforming real estate analytics. Utilizing the "Housing & Ocean Proximity" dataset, it aims to forecast housing prices, analyze crime rates, and identify areas ripe for revamping. The study underscores the importance of integrating data on house characteristics and neighborhood dynamics to predict market trends. By harnessing predictive modeling, the research seeks

to offer valuable insights for stakeholders in real estate, urban planning, and policy-making, highlighting the synergy between data science and real estate market analysis.

Related Work:

Research Paper Citation

Paper 1: Endogenous Gentrification and Housing Price Dynamics

The paper discusses the variation in house price growth in different neighborhoods within a city during overall housing booms. It introduces a model linking these price movements and neighborhood gentrification following citywide housing demand shocks. A key concept is the preference for living near wealthier neighbors, leading to income-based segregation. Higher-income residents move into adjacent poorer areas during demand spikes, driving up prices and displacing original residents, a process termed "endogenous gentrification." The paper provides empirical evidence supporting this model, using various data sets and city-level demand shock analysis.

<u>Paper 2:</u> Stratifying and predicting patterns of neighborhood change and gentrification: An urban analytics approach

This paper tackles the complexity of identifying and differentiating gentrification from other types of neighborhood changes in cities, using London as a case study. It employs a novel urban analytics approach, integrating diverse datasets on population, house prices, and development. The study uses data reduction and classification methods, followed by machine learning, to analyze and predict gentrification trends.

Data Definitions

It is a dataset that contains information about houses and their proximity to the ocean. The dataset contains the following data features:

- Address: The address of the house.
- Latitude and longitude: The latitude and longitude of the house.
- **Distance to ocean:** The distance from the house to the ocean.
- **Type of house:** The type of house (e.g., single-family home, apartment, condo).
- **Number of bedrooms:** The number of bedrooms in the house.
- **Number of bathrooms:** The number of bathrooms in the house.
- **Square footage:** The square footage of the house.
- **Price:** The price of the house.

The data comes from a variety of sources, such as public records, real estate listings, and surveys.

Sources of housing ocean proximity data that I have got access to use, is from my previous CMU Project for an HCII elective. It is varied and includes coastal survey databases managed by governmental agencies, which provide detailed information about properties near the ocean. Real estate websites often feature filters and maps highlighting oceanfront or ocean-close properties. GIS data from national and local sources can offer precise mapping of property locations in relation to the coastline.

Baseline Performance

=== Run information ===

weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 Scheme: -W1-K "weka.classifiers.functions.supportVector.PolyKernel -E1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4" Relation: h_700-weka.filters.unsupervised.attribute.Remove-R11 Instances: 700 Attributes: 10 longitude latitude housing_median_age total_rooms total bedrooms population households median_income median house value ocean_proximity Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used: Linear Kernel: K(x,y) = <x,y>

Classifier for classes: <1H OCEAN, INLAND

BinarySMO

Machine linear: showing attribute weights, not support vectors.

4.8876 * (normalized) longitude

- + 6.3828 * (normalized) latitude
- + -1.0439 * (normalized) housing_median_age
- + 2.4531 * (normalized) total_rooms
- + 0.3129 * (normalized) total_bedrooms
- + -1.4328 * (normalized) population
- + -0.5887 * (normalized) households
- + 1.1885 * (normalized) median_income
- + -4.9497 * (normalized) median_house_value
- 2.9463

Number of kernel evaluations: 23809 (69.374% cached)

Time taken to build model: 0.03 seconds

=== Stratified cross-validation === === Summary ===

Correctly Classified Instance	es 644	92	%
Incorrectly Classified Instan	ces 56	8	%
Kappa statistic	0.8334		
Mean absolute error	0.08		
Root mean squared error	0.2828		
Relative absolute error	16.4342 %		
Root relative squared error	57.3328 %		
Total Number of Instances	700		

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC **ROC** Area PRC Area Class 0.966 0.143 0.903 0.966 0.933 0.836 0.911 0.892 <1H OCEAN 0.857 0.034 0.947 0.857 0.900 0.836 0.911 0.871 INLAND Weighted Avg. 0.920 0.098 0.922 0.920 0.919 0.836 0.911 0.884

=== Confusion Matrix ===

a b <-- classified as 393 14 | a = <1H OCEAN 42 251 | b = INLAND

The model was trained using a dataset with 700 instances and 10 attributes, including geographical (longitude, latitude), demographic (population, households), and economic (median income, median house value) factors, among others. The goal is to classify instances into two classes: <1H OCEAN and INLAND, probably based on geographical and other related attributes.

Key points from the output:

- **SMO Algorithm:** It utilized a linear kernel, indicated by the formula showing attribute weights. Linear kernels are simpler and faster but might be less powerful for complex datasets.
- Model Performance: The model achieved a high accuracy of 92% on 10-fold cross-validation. This suggests that it was generally effective in classifying instances correctly.
- **Evaluation Metrics:** Various metrics like Kappa statistic, mean absolute error, and others are provided. The Kappa statistic of 0.8334 indicates a strong agreement between predicted and actual classifications. The errors (mean absolute and root mean squared) are relatively low, indicating good predictive power.
- Class-wise Performance: The detailed accuracy by class shows that the model performed better in predicting <1H OCEAN compared to INLAND, as seen in the TP (True Positive) Rate and Precision.
- **Confusion Matrix:** It gives a clear picture of the model's performance in terms of false positives and false negatives. Most errors seem to be in falsely classifying INLAND instances as <1H OCEAN.

Error Analysis Process

The error analysis process involves several steps:

The **confusion matrix** gives a breakdown of the model's performance, illustrating where it makes correct predictions and where it errs. It is evident that the model has certain limitations, such as the potential misclassification of the INLAND class.

Feature Impact Assessment : Using the 'Feature Weight' information to determine which features have the most significant impact on predictions.

Features with very high weights may be dominating the model, potentially leading to overfitting.

Using **Misclassification Review** for examining specific instances of false positives and false negatives to understand the context of these errors.

Problems Identified:

- **Imbalance in Prediction -** The model might be biased towards predicting one class over the other, as indicated by the disparity in false positives and false negatives.
- Feature Influence Some features might be disproportionately influencing the model's predictions, as indicated by the feature weight analysis. Overfitting - Given the relatively high accuracy but lower Kappa, the model might be overfitting the training data.
- Potential Improvements To enhance the model, the following strategies could be introduced: Feature Engineering - Enhance the feature set by adding new features or transforming existing ones, like creating interaction terms between latitude and longitude, or categorizing continuous variables.
- Parameter Tuning Use regularization more effectively to prevent overfitting. This could be done by adjusting the regularization strength (L1, L2) or using a dual approach. Class Weight Adjustment - If the classes are imbalanced, adjust the class weights in the Logistic Regression to make the model more sensitive to the minority class. Cross-validation - Implement a more robust cross-validation scheme to ensure the model generalizes well.

A structured **evaluation experiment** would involve the following steps: Implement Changes - Introduce the improvements to the feature set and model configuration. Retrain Model - Use the modified features and parameters to retrain the Logistic Regression model. Validate Improvements -Assess performance on a validation set or through cross-validation, focusing on Kappa, Accuracy, and the confusion matrix. Compare Performance -Evaluate whether changes have led to a statistical improvement over the baseline model. For parameter tuning, tools like CVParameterSelection would be employed, focusing on numerical parameters like the regularization coefficient. The chosen parameters would be those that maximize the cross-validated performance metrics.

This approach to error analysis and model improvement is iterative and data-driven, relying on a deep understanding of the model's current limitations and the data's underlying patterns. By addressing the identified issues and meticulously evaluating the impact of the changes, the model's predictive power can be enhanced.

•		Light	Side		
	Extract Features	Restructure Data Build Models	Explore Results	Compare Models Predict	Labels
CSV Files:	Ľ	Feature Extractor Plugins:		Configure Column Fe	eatures
h_700.csv DOCUMENT_LIST Class: Ocean_ Type: NOMIN Cext Fields: Sr. No. households housing_med latitude Differentiate	700 .proximity (IAL (Jian_age			Column Name Feature Ty Filenames NOMINAL Sr. No. NUMERIC housing_me NUMERIC latitude NUMERIC latitude NUMERIC median_hou NUMERIC population NUMERIC population NUMERIC total_bedroo NUMERIC total_rooms NUMERIC	pe Values [/Users/ashishm 700 unique values 493 unique values 283 unique values 332 unique values 573 unique values 664 unique values 621 unique values 651 unique values
Extract	Name: columns	_1 Rare Threshold:	5		
eature Table:	E	• Evaluations to Display:		Features in Table:	ka
columns FEATURE_TABLE > Documents: > Feature Plug > Feature Tab 9 features	gins: columns	Target: <1H OCEAN		Search: Feature households_column housing_median_age_column latitude_column longitude_column median_house_value_column median income column	

Image: A Screenshot of Feature Extraction using Column Features :Text

fields for extraction include demographics and location data. Feature extraction plugins for columns and English parsing are selected, and the

interface lists numerical features with their unique value counts. The feature table 'columns' is ready for analysis targeting the '<1H OCEAN' class, with various performance metrics indicated but not their specific values.

Extract Features	Restructure Data	Build Models	Explore Res	ults Com	npare Models	Predict Labels		
COLUMNS	Naive Bay Logistic F Linear Re Support V Decision	res Regression gression /ector Machines Trees				L2 Regularizat	tion	essio
1	Cross-Va	llidation Test Set ation N	 Random By Annotation Sr. No. By File umber of Folds Auto Manual: 10 	on: ::	G May			
Name: logit_colu	umns_3	Feature Se						
E	Model Evalua	tion Metrics:	•		Model Confu	sion Matrix:		
_700.csv ns: columns :: columns	Accuracy Kappa	0.8	8957		Act \ Pred <1H OCEAN INLAND	<1H OCEAN 381 47	INLAND 26 246	
	Columns proximity I Name: Ogit_colu 2 Columns proximity I Columns proximity I Columns	Learning Plug Naive Bay Logistic F Linear Re Support N Decision Weka (All Evaluation Op Cross-Va Supplied No Evalu No Evalu No Evalu Model Evalua 2 C M Metric Accuracy Kappa	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machines Decision Trees Weka (All) Evaluation Options: For Cross-Validation Cross-Validation Supplied Test Set No Evaluation No Evaluation Name: Ogit_columns_3 Feature Set Model Evaluation Metrics: 2 S Construction Construc	Image: Solution of the second sec	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machines Decision Trees Weka (All) proximity I Evaluation Options: Fold Assignment: O Cross-Validation Supplied Test Set By File Number of Folds: Auto Manual: 10 2 Supplied Test Set No Evaluation Supplied Test Set No Evaluation Supplied Test Set No Evaluation Sr. No. By File Number of Folds: Auto Model Evaluation Metrics: 2 Xappa Name: logit_columns_3 Feature Selection	Image: Superior Vector Machines 2 00.csv 1 2 0 1 2 0 1 <	Image: Configure Logis Image: Configure Logis Image: Columns L700.csv Is: columns Decisit Regression Decisit Regression Support Vector Machines Decision Trees Weka (All) Evaluation Options: Fold Assignment: O Cross-Validation Supplied Test Set By File Number of Folds: Auto Manual: 10 2 2 Model Evaluation Metrics: Model Evaluation Metrics: Model Evaluation Metrics: Model Confusion Matrix: Metric Value Accurracy Accu	Image: Section Section Image: Section Sectio

Image: A Screenshot of Build Models using Logistic regression with Cross Validation to check the performance of the dataset h_700.csv

Here, Cross-validation is set to 'Random' with an automatic fold assignment. The Model Evaluation Metrics show **an accuracy of 0.8957 and a kappa of 0.7836.** The confusion matrix for the predictions has 381 correct for '<1H OCEAN' and 246 correct for 'INLAND', with 26 and 47 instances misclassified, respectively.

		Lig	htSide		
	Extract Features	Restructure Data Build Models	Explore Results Compar	e Models Predict Labels	
Feature Tables: columns FEATURE_TABLE > Documents: h > Feature Plugin > Feature Table 9 features Class: ocean_ Type: nomina	proximity	Learning Plugin: Naive Bayes Logistic Regression Linear Regression Support Vector Machines Decision Trees Weka (All) Evaluation Options: Cross-Validation Supplied Test Set No Evaluation	Test Set (CSV): h_300.csv DOCUMENT_LIST Documents: h_300.csv Files: Instances: 300 Text Columns: 0		ession
🖄 Train	Name: logit_colum	nns_4	election		
Trained Models: logitcolumns_ TRAINED_MODEL > Documents: h > = Feature Plugi > = Feature Table > = Learning Plug	n_700.csv ns: columns	Accuracy 0.	alue A	Indel Confusion Matrix: Act \ Pred <1H OCEAN INLAND C1H OCEAN 160 12 NLAND 21 107	

Image: A Screenshot of Build Models using Logistic regression with Supplied Test Set to check the performance of the dataset h_300.csv

Logistic Regression with L2 Regularization is selected and evaluated on a test set 'h_300.csv'. Metrics show an **accuracy of 0.89 and a kappa of 0.7731.** The confusion matrix details predictions for two classes: '<1H OCEAN' (160 correct, 12 incorrect) and 'INLAND' (107 correct, 21 incorrect)

Conclusion: The cross-validations performance is poor than the Supplied Test Performance.

Extract Features Rest	tructure Data Build Models Explore Results Compare Models Predict Labels
Highlight:	Cell Highlight: 😽 Features in Table:
logit_columns_3 😌 📄 🗙	Act \ Pred <1H OCEAN
TRAINED_MODEL Documents: h_700.csv Feature Plugins: columns Feature Table: columns Learning Plugin: Logistic Regression Validation: h_300.csv Validation: h_300.csv Trained Model: logit_columns_3 Kappa: 0.773 Accuracy: 0.890	INLAND 21 107 Feature Average C Horizonta Vertical A Feature Iatitu 35.8329 0.9438 1.3072 -0.6566 medi 4.2532 1.3562 0.2021 -0.2736 longit -119.4229 0.2546 0.6055 -0.156 total 430.9048 104.1139 87.6077 -0.0016 total 420.9048 104.1139 87.6077 -0.0016 total 430.8571 62.1896 94.6616 0.0062 popul 1138.23 250.7152 484.9119 0.001 hous 430.8571 62.1896 94.6616 0.0062 housi 27.8095 5.146 1.9842 0.0375 Feature Confusion Ranking Vertical Absolute Difference Vertical Absolute Difference Vertical Absolute Difference Horizontal Difference Vertical Absolute Difference Vertical Absolute Difference Vertical Absolute Difference
Exploration Plugin: Highlighted Feature I	
Average Cell Value	Horizontal Absolute Difference Vertical Absolute Difference
Model Confusion Matrix: Act \ Pred <1H OCEAN	Model Confusion Matrix: Image: Confusion

Image: A Screenshot of Explore Results for error analysis showing Average Cell Value, Horizontal Absolute Difference & Vertical Absolute Difference

Description: Here, Evaluation metrics for the Logistic Regression model is with a kappa of 0.773 and accuracy of 0.890. The confusion matrix indicates 160 correct predictions for '<1H OCEAN' and 107 for 'INLAND', with a few misclassifications. The interface also presents detailed feature impact analyses like average cell values and absolute differences, relating to features such as latitude and median_house_value.

Extract Features Res	tructure Data Build Models Explore Results Compare Models Predict Labels	
Highlight:	Cell Highlight: 😽 Features in Table:	
logit_columns_3 😌 📔 🗙	Act \ Pred <1H OCEAN	
TRAINED_MODEL Documents: h_700.csv Feature Plugins: columns Feature Table: columns Learning Plugin: Logistic Regression Validation: h_300.csv Validation: h_300.csv Trained Model: logit_columns_3 Kappa: 0.773 Accuracy: 0.890	INLAND 21 107 Feature Average Horizont Vertical A. housing 6.3.0833 6.7104 0.4198 househol 642.75 117.2313 149.7033 populati 1681.8 58.6833 292.8801 median 142700 111010.7 43245.7. total_roo 3664.9 883.1292 951.3746 total_roo 156.0736 130.5644 130.5644	-0.0375 3 -0.0062 1 -0.001 0 5 0.0014
Exploration Plugin: Highlighted Feature	Details 📀 Calculating row and colu	umn values
Horizontal Absolute Difference	Vertical Absolute Difference Feature Weight	
	Model Confusion Matrix:	
Act \ Pred <1H OCEAN	Act \ Pred <1H OCEAN	

Image: A Screenshot of Explore Results for error analysis showing Feature Weight

Description: Same as for last Image

Extract Features	Restri	ucture Data	Bullo	d Models	Expl	ore Resul		ompare M	nouers	Predict	Labels		
Model to Apply:		Selected	Dataset	t: h_700.	csv (oce	an_proxi	mity_pr	ediction)				
logitcolumns_3 😔 📋	×	Sr. No.			latitude	5						total_b	. total_r
	<u> </u>	301	444	29	33.86					<1H		475	2787
TRAINED_MODEL		302	336	19	38.11	-122.6	201600	3.8068	<1H	INLAND	873	328	1752
> 🕅 Documents: h 700.csv		303	832	38	33.85					<1H		862	3596
> R Feature Plugins: columns		304	356	39	33.97					<1H		380	1346
> Eature Table: columns		305	334	36	34.09					<1H		392	3129
		306	706	13	38.2					INLAND		847	4110
Learning Plugin: Logistic Regressio	n	307	411	13	36.58					INLAND		405	1788
📷 Validation: h_300.csv		308	853	28	34.1	-117				<1H		871	4086
V 🖄 Trained Model: logit_columns_3		309	297	35	33.91					<1H		302	1092
Kappa: 0.773		310	546	52	34.02					<1H		587	2511
Accuracy: 0.890		311	338	16	38.37					INLAND		331	2495
Accuracy. 0.050		312	254	15	33.76					<1H		297	851
		313	200	25	33.99					<1H		210	1348
		314	372	21	36.13					INLAND		376	2271
		315	708	19	37.4					<1H		764	4043
		316	379	45	34.04					<1H		417	1767
		317	474	7	37.33					INLAND		621	3389
Copy Validation Results to Test Data	a	318	314	44	33.95					<1H		338	1812
		319	1387	11	35.32					INLAND		1455	7035
New Data: 🔬		320	389	8	38.02	-121				INLAND		392	1868
		321 322	375 788	41 36	36.76 34.09					INLAND		399 833	1973 3503
h_700.csv (ocean_proximit 😒	X	322	788 75	25	34.09					<1H		833 74	434
		323	75 428	25 18								74 715	3903
DOCUMENT_LIST		324	428	4	34.28 37.76					INLAND		1439	6875
Documents: h_700.csv (ocean_pro:	ximity	325	171	4 52	37.98					< IH INLAND		1439	941
> Files:		320	439	52 6	34.14					<1H		506	1727
Instances: 700		328	439	4	34.14					< IR		510	3078
		329	616	4 46	34.00	-121.5				INLAND		645	2646
Text Columns: 0		330	3478	5	34.26					INLAND		3521	25187
		331	401	16	33.91					<1H		423	2889
		332	550	29	33.91					<1H		423 582	2784
		333	403	23	33.79					<1H		430	2663
		334	523	22	33.65					<1H		527	3592
		335	333	18	34.15					<1H		420	1880
		336	526	42	37.34					<1H		524	2101
		337	813	13	33.88					<1H		849	3239
													(5)
Predict New Column Name	e: oxir	nity_predic	tion	🔺 🗌	Show L	abel Dist.	tributior	1 🗌 C	Overwrite	Column	S		

Image: A Screenshot of Predict Labels for error analysis

Description: The model's evaluation metrics are listed as a kappa of 0.773 and an accuracy of 0.890. The columns include various housing-related features, and the actual versus predicted 'ocean_proximity' classifications. Rows of data entries are visible, with some showing discrepancies between the actual and predicted labels.

	A	в	С	D	E	F	G	н	1	J	К	L	
1	Sr. No. =	households -	housing_media -	latitude -	longitude -	median_house			ocean_proximity_prediction T	population -	total_bedrooms -	total_rooms =	
411	308	853	28	34.1	-117.76	202200	2.621		<1H OCEAN	1973	871	4086	
422	337	813	13	33.88	-117.59	107000	2.6111	INLAND	<1H OCEAN	2751	849	3239	
432	359	613	2	34.38	-118.61	329500	6.6916	INLAND	<1H OCEAN	1787	883	5989	
438	385	338	33		-117.97	143900	2.9712	INLAND	<1H OCEAN	1600	316	1558	
458	443	395	7	37.81	-121.91	500001	13.1499		<1H OCEAN	1216	416	3477	
459	445	183	37	34.08	-118.03	159200	3.25		<1H OCEAN	726	179	775	
469	470	634	36	34.14	-117.86	235300	3.1905		<1H OCEAN	1484	667	3097	
470	471	14		38.63	-121.28	350000	10.2264		<1H OCEAN	30	16	120	
473	476	450		33.93	-117.48	122000	2.6776		<1H OCEAN	1564	459	2191	
475	478	436		34.06	-117.42	143100		INLAND	<1H OCEAN	1305	495	2532	
475	470	484	21	34.00	-117.64	102500		INLAND	<1H OCEAN	2556	507	1801	
484	503	551	7	34.04	-117.04	225000	1.4007		<1H OCEAN	1587	647	2461	
404	528	501	26		-117.6	153100	3.1859		<1H OCEAN	1921	575	2925	
491	530	14			-117.12	162500		INLAND	<1H OCEAN	46	14	2925	
	530	14		34.1	-116.23	76900			<1H OCEAN	46 3779		4981	
499			32							1862	1326	2643	
526	601	478			-117.78	177200		INLAND	<1H OCEAN		516		
534	623	405			-118.01	166300	3.4609		<1H OCEAN	1375	412	2120	
537	630	190	11	33.67	-117.07	145800	2.375		<1H OCEAN	557 2258	187 746	939	
546	652	672			-117.51	124700	0.2001		<1H OCEAN			3791	
559	677	233			-117.73	118100		INLAND	<1H OCEAN	809	243	975	
592	763	55			-118.09	237500	2.2321		<1H OCEAN	150	74	271	
604	790	138		34.06	-117.65	112500	2.0893		<1H OCEAN	349	130	465	
633	865	191	35	34.08	-117.99	134800	2.8906		<1H OCEAN	954	207	1032	
640	879	905		33.51	-116.01	66400	1.7344		<1H OCEAN	4042	958	2985	
649	897	431	21	37.94	-121.95	285400	6.8642		<1H OCEAN	1318	411	3153	
654	908	352	24	34.14	-117.98	148000	3.0417	INLAND	<1H OCEAN	1329	388	1596	
657	914	1015	20	38.41	-122.4	267600	2.5685		<1H OCEAN	1725	1015	4867	
671	944	412	36	34.11	-118.06	239500	2.7656	INLAND	<1H OCEAN	914	485	2178	
677													
011	962	858			-117.65	149100	3.449		<1H OCEAN	2373	903	4420	
677 691	962	858 348			-117.65 -117.73			INLAND INLAND	<1H OCEAN <1H OCEAN	2373 1459	903 378	4420 1967	
	984	348	42	34.03	-117.73	118100	3.0375	INLAND	<1H OCEAN	1459	378		
691	984 A	348 B	42 C	34.03 D	-117.73 E	118100 F	3.0375 G	H	<1H OCEAN	1459 J	378 K	1967 L	
691 1	984 A Sr. No. -	B households ;	C housing_media -	D latitude ,	-117.73 E Iongitude ,	118100 F median_house_ پ	G median_incom	H ocean_proximity	<1H OCEAN	J population =	378 K total_bedrooms ب	L total_rooms =	
691 1 3	984 A Sr. No	B households ऱ् 336	C housing_media = 19	عد 34.03 D Iatitude ج	-117.73 E Iongitude , -122.6	F median_house_ ب 201600	G median_incom(▼ 3.8068	H ocean_proximity Y <1H OCEAN	<1H OCEAN	ل پ population ج 873	378 K total_bedrooms ب 328	1967 L total_rooms ب 1752	
691 1 3 18	A Sr. No. ج 302 330	B households 336 3478	242 C housing_media ج 19 5	D latitude − 38.11 34.26	-117.73 E longitude — -122.6 -118.9	F median_house_ چ 201600 321300	G median_incom، ۲ 3.8068 6.9712	H ocean_proximity Y <1H OCEAN <1H OCEAN	<1H OCEAN I ocean_proximity_prediction ¥ INLAND INLAND	ل پ population چ 873 11956	378 K total_bedrooms ب 328 3521	1967 L total_rooms = 1752 25187	
691 1 3 18 25	A Sr. No. = 302 330 338	B households	42 C housing_media ج 19 5 2	D latitude 38.11 34.26 33.65	-117.73 E Iongitude - -122.6 -118.9 -117.59	F median_house_ ╤ 201600 321300 151900	G median_incom ▼ 3.8068 6.9712 4.5022	H ocean_proximity Y <1H OCEAN <1H OCEAN <1H OCEAN	I OCEAN	J population ₹ 873 11956 2332	378 K total_bedrooms , 328 3521 1193	1967 L total_rooms 	
691 1 3 18 25 27	A Sr. No. 〒 302 330 338 340	B households ╤ 336 3478 1073 401	C housing_media - 19 5 2 2 26	24.03 D 38.11 34.26 33.65 39.13	-117.73 E -122.6 -118.9 -117.59 -123.2	F median_house_ ╤ 201600 321300 151900 84400	G median_incom ♥ 3.8068 6.9712 4.5022 1.375	H ocean_proximity Y <1H OCEAN <1H OCEAN <1H OCEAN <1H OCEAN	I I OCEAN I OCEAN INLAND INLAND INLAND INLAND	J population 873 11956 2332 1065	378 K total_bedrooms - 328 3521 1193 417	1967 L total_rooms ₹ 25187 4860 1474	
691 1 3 18 25 27 41	A Sr. No. 〒 302 330 338 340 366	B households 336 3478 1073 401 584	42 C housing_media → 19 5 2 26 17	D latitude 38.11 34.26 33.65 39.13 39.18	-117.73 E -122.6 -118.9 -117.59 -123.2 -123.21	F median_house_ ╤ 201600 321300 151900 84400 142100	G median_incom ♥ 3.8068 6.9712 4.5022 1.375 2.6275	INLAND H cean_proximity <1H OCEAN	I OCEAN I OCEAN INLAND INLAND INLAND INLAND INLAND INLAND	J population 873 11956 2332 1065 1501	378 K total_bedrooms 	L total_rooms 25187 4860 1474 2772	
691 1 3 18 25 27 41 46	A Sr. No. 302 330 338 340 366 371	B households 3360 3478 1073 401 584 492	C housing_media = 19 5 2 2 26 17 18	D latitude ╤ 38.11 34.26 33.65 39.13 39.18 34.93	-117.73 E -122.6 -118.9 -117.59 -123.2 -123.21 -123.21 -120.44	F median.house. ⇒ 201600 321300 151900 84400 142100 67500	G median_incom(¥ 3.8068 6.9712 4.5022 1.375 2.6275 2.1509	INLAND H ocean.proximity <1H OCEAN	I OCEAN	J population	K total_bedrooms ≠ 328 3521 1193 417 576 558	L total_rooms 〒 225187 4860 1474 2772 2098	
691 1 3 18 25 27 41 46 68	A Sr. No. 3002 3300 338 340 386 371 401	B households 〒 3366 3478 1073 401 584 492 341	C housing.media = 19 5 26 26 17 17 16 15	D 1atitude 〒 38.11 34.26 33.65 39.13 39.13 39.13 39.13 39.13 33.65 39.13	-117.73 E -122.6 -118.9 -117.59 -117.59 -123.21 -123.21 -120.44 -121.8	F median house. ╤ 201600 321300 151900 84400 142100 67500 164500	G median_incom ¥ 3.8068 6.9712 4.5025 1.375 2.21509 4.5045	INLAND H cean_proximity CHOCEAN CHOCEAN	<th ocean<br="">I ocean_proximity_prediction T INLAND INLAND INLAND INLAND INLAND INLAND</th>	I ocean_proximity_prediction T INLAND INLAND INLAND INLAND INLAND INLAND	J population = 873 11956 2332 1065 1501 1252 1252 1257	378 K total.bedroomt ~ 328 3521 1193 417 576 558 378	ل total.rooms ح الم الم الم ا المح الم الم الم الم الم الم الم الم
691 1 3 18 25 27 41 46 68 105	A - Sr. No. - 302 - 330 - 338 - 340 - 366 - 371 - 401 -	B households 336 3378 1073 401 584 492 341 396	C housing_media ╤ 19 5 2 2 2 6 17 16 15 20	D attude → 38.11 34.26 33.65 39.13 39.18 34.93 34.93 37.31 38.25	-117.73 E -122.6 -118.9 -117.59 -123.2 -123.2 -123.21 -120.44 -121.8 -122.62	F median.house. = 201600 321300 151900 84400 142100 67500 164500 189100	G median_incom ▼ 3.8066 6.9712 4.5022 1.375 2.6275 2.1509 4.5045 2.875	INLAND H Ocean_proximity CH OCEAN <1H OCEAN	I OCEAN	1459 population → 873 11956 2332 1065 1501 1252 1277 826	K total_bedrooms 328 3521 1193 417 558 378	L total_rooms = 25187 4860 1474 2772 2098 1807 1888	
691 1 3 18 25 27 41 46 68 105 144	A Sr. No. 302 330 338 340 366 371 401 462 523	B households ⇒ 3366 3478 1073 401 584 402 341 396 915	C housing.media = 19 5 22 6 117 16 16 15 20 9 9	D ■ 38.111 34.26 33.65 39.13 39.18 34.93 37.31 33.25 38.42 38.42	-117.73 E -122.6 -118.9 -117.59 -123.21 -123.21 -123.21 -121.48 -122.62 -122.79	F median house, ∓ 201600 321300 151900 84400 142100 67500 164500 188100 188500	G median_incom ¥ 3.8066 6.8712 4.5022 1.375 2.6275 2.1509 4.5045 2.875 5.038	INLAND H ocean, proximity <1H OCEAN	I OCEAN	1459 population ₹ 873 11956 2332 1065 1601 1252 1277 826 2581	378 K total.bedroom 328 3521 1193 417 576 558 378 378 411 885	L L L L L L L L L L L L L L L L L L L	
691 1 3 18 25 27 41 46 68 105 144 152	984 A Sr. No. 〒 302 330 338 340 386 371 401 462 523 542	B households ⊽ 336 3478 1073 401 584 492 341 396 915 1158	C housing_media = 19 5 2 2 6 6 19 7 7 16 15 20 0 9 9 14	Jatitude マ Iatitude マ 34.03 38.11 34.26 33.65 39.13 39.13 39.33 39.73 37.31 38.25 38.42 38.42	-117.73 E longitude ⊽ -122.6 -118.9 -117.59 -123.21 -120.44 -121.8 -122.82 -122.79 -122.79	118100 F median.bus. ₹ 201600 321300 151900 142100 67500 144200 144500 148500 1185600 1135400	G median_incoms ▼ 3.8068 6.9712 4.5502 1.375 2.8275 2.1509 4.5045 2.875 5.5038 2.0651	INLAND H ocean, proximity <1H OCEAN	I OCEAN	1459 population 〒 873 11968 2332 1066 1501 1252 1277 8268 2581 2323	K total_bedrooms ₹ 328 3521 1193 417 558 378 411 885 1298	L total_rooms = 25187 4860 1474 2772 2098 1807 1888 4967 4042	
691 1 3 18 25 27 41 46 68 105 144 152 218	984 A Sr. No. ⇒ 302 330 338 340 396 371 401 462 623 642 642 655	B households 336 3478 1073 401 584 492 396 915 1158 157	C housing_media = 19 5 2 2 2 2 2 2 2 2 2 3 1 4 3 3	Jatitude	-117.73	118100 F 201600 321300 84400 142100 164500 189100 189100 135400 36700	G median_incoms ¥ 3.8068 6.87712 4.5022 1.375 2.8275 2.8509 4.5045 2.875 5.038 2.0851 1.0488	INLAND H ocean, proximity <1H OCEAN	I OCEAN	1459 population 〒 873 11956 2332 1065 1501 1252 1277 826 2581 2333 425	378 K total_bedroomi ╤ 328 3521 1193 417 576 558 378 417 855 1288 228	ل total_rooms ح 25187 4860 1474 2772 2098 1807 1888 4967 4042 938	
691 1 3 18 25 27 41 46 68 105 144 152	984 A 302 330 338 340 366 371 401 462 523 542 655 683	B households 〒 336 3478 1073 401 584 492 341 492 341 1158 1158 157 414	C housing.media 7 19 5 22 26 17 16 15 20 9 14 31 31 16	Jatitude ▼ 38.11 34.26 33.65 39.13 39.18 34.03 37.31 39.18 37.31 38.42 38.42 38.42 41.32 39.44	-117.73 E longitude -122.6 -118.9 -123.21 -123.21 -123.21 -122.42 -122.42 -122.42 -122.79 -122.73 -122.85 -122.85 -122.85 -122.85 -122.85	118100 F 201600 321300 151900 84400 142100 164500 189100 185600 138400 138500 1354000 1354000 1354000 1354000000000000000000000000000000000000	G median_incoms ¥ 3.8068 6.8712 4.5022 1.375 2.8275 2.1509 4.5045 2.875 5.038 2.0651 1.0488 3.2171	INLAND H ocean_proximity <1H OCEAN	I OCEAN	1459 population ₹ 873 11956 2332 1065 1501 1252 1277 826 2581 2283 425 1177	378 K total_bedroomt 328 3521 1193 376 576 558 376 411 885 1298 423 423	1967 L total_rooms 〒 1752 25187 4860 1474 2772 2008 1807 1808 4967 4062 938 2017	
691 1 3 18 25 27 41 46 68 105 144 152 218	A A Sr. No. 302 330 338 340 366 371 401 462 523 542 665 663 684	B households	C housing.media = 19 5 22 6 26 117 16 15 20 9 14 30 16 37	Jatitude ▼ 38.11 34.68 33.65 39.13 39.18 34.93 37.31 39.18 34.63 37.31 38.42 38.42 38.42 38.46 39.44 38.49	-117.73 E longitude -122.6 -118.9 -123.21 -123.21 -123.21 -122.42 -122.42 -122.73 -122.73 -122.73 -122.85 -122.79 -122.73 -122.85 -123.79 -133.79	118100 F median. house. ₹ 201600 321300 151900 84400 142100 142100 1485000 185600 135400 135400 115200 116200 146500	G median_incomi ¥ 3.8068 6.9712 4.5022 1.375 2.8275 2.1509 4.5045 2.875 5.038 2.0651 1.0488 3.2171 3.8343	INLAND H ocean, proximity <1H OCEAN	I OCEAN I OCEAN I OCEAN INLAND	1459 population 〒 873 11956 2332 1065 1601 1252 1277 826 2323 425 1177 1137	378 K total.bedroomt ≠ 328 3521 1193 417 576 558 378 411 885 1288 238 423 423 519	1967 L total.roms 〒 25187 4860 1474 2772 2098 1807 1808 4967 4042 938 4967 4042 938 2017 2468	
691 1 3 18 25 27 41 46 68 68 105 144 152 218 233	A A Sr. No. 302 330 340 386 371 401 462 523 655 663 684 738	B households 3376 3376 3478 401 584 402 340 584 402 341 396 9155 1158 1157 414 474 474 320	C housing_media = 19 5 2 2 2 2 6 17 16 15 20 9 9 114 31 16 37 30	Jatitude	-117.73 E Iongitude -122.6 -118.9 -123.2 -123.2 -123.2 -123.2 -122.73 -122.73 -122.73 -122.73 -122.73 -122.79 -122.79 -122.79 -122.79 -122.79 -122.79 -122.79 -122.85 -123.79 -123.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.75 -123.85 -123.75 -123.85 -123.75 -123.85 -123.75 -123.85 -123.75	118100 F 201600 321300 151900 84400 142100 67500 164500 189100 189100 135400 135400 1136000 146500 146500	G median_incomx ¥ 3.8068 6.8712 4.5022 1.375 2.1509 4.5045 2.875 5.038 2.0651 1.0488 3.2171 1.33343 5.0288	INLAND H ocean_proximity <1H OCEAN	I OCEAN I OCEAN I OCEAN INLAND	1459 population	578 K total_bedroomi ╤ 328 3321 1193 417 556 358 411 885 1298 423 519 423 519 342	لو total_room ت 25187 4860 1474 2772 2098 1807 1807 1808 4967 4042 938 2017 2466 1973	
691 1 3 18 25 27 41 46 68 105 144 152 218 233 234	984 A Sr. No. ⇒ 302 330 338 340 386 340 386 340 386 351 401 462 523 542 655 663 684 738 750 750	B households 〒 336 3478 1073 401 584 402 341 396 915 1158 1158 1158 1157 414 474 474 474	C housing_media 7 19 5 2 2 6 17 16 15 20 9 14 31 16 37 30 0 22	Jatitude ▼ 38.11 34.26 38.81 39.13 39.13 39.13 39.13 39.13 33.7.31 38.25 38.46 41.32 39.44 38.49 34.99 41.3	-117.73 E longitude -122.6 -118.9 -123.2 -123.21 -120.44 -121.8 -122.82 -122.79 -122.73 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.85 -123.79 -122.81 -122.85 -123.79 -122.81 -123.85 -	118100 F 201600 321300 151900 84400 142100 164500 189100 1885000 1385000 38700 118200 1146200 1146500 116200 116200 158000 62700	G median_incoms ¥ 3.8068 6.8712 4.5042 2.1509 4.5045 2.875 5.038 2.0651 1.0486 3.2171 3.343 5.0286 1.8065	INLAND H ocean_proximity <1H OCEAN	I OCEAN I OCEAN I OCEAN INLAND	1459 population - 873 11966 2332 1065 1501 1252 1277 826 22581 2323 425 1177 1137 999 999 686	x78 K total_bedroomt 328 3521 1193 417 576 558 378 4111 885 1298 423 1298 423 519 3424 372	1967 L total_rooms ₹ 1752 25187 4860 1474 2772 2098 1807 1807 1807 1807 1808 4967 1975 2466 1975 2466 1975 1580	
691 1 3 18 25 27 41 46 68 105 144 152 218 233 234 267	984 A 302 330 338 340 366 371 401 462 553 643 738 750 762	B households 3376 3376 3478 401 584 402 340 584 402 341 396 9155 1158 1157 414 474 474 320	C housing.media = 19 5 2 2 2 2 2 2 2 3 4 19 10 10 10 10 10 10 10 10 10 10 10 10 10	Jatitude マ 38.11 34.26 33.65 39.13 39.18 34.93 37.31 39.18 38.42 38.46 41.32 39.44 38.49 34.49 34.43 34.49 38.44 38.49 34.43 34.49 34.43 34.49 34.43 34.89 41.3 32.83	-117.73 E Iongitude -122.6 -118.9 -123.2 -123.2 -123.2 -123.2 -122.73 -122.73 -122.73 -122.73 -122.73 -122.79 -122.79 -122.79 -122.79 -122.79 -122.79 -122.79 -122.85 -123.79 -123.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.75 -123.85 -123.75 -123.85 -123.75 -123.85 -123.75 -123.85 -123.75	118100 F median. house. ₹ 201600 321300 151900 84400 142100 164500 188100 188500 138400 138400 118200 146500 146500 146500 146500 148500 1	G median_incomx ¥ 3.8068 6.8712 4.5022 1.375 2.1509 4.5045 2.875 5.038 2.0651 1.0488 3.2171 1.33343 5.0288	INLAND H ocean_proximity <1H OCEAN	I OCEAN I OCEAN I OCEAN INLAND	1459 population	578 K total_bedroomi ╤ 328 3321 1193 417 556 358 411 885 1298 423 519 423 519 342	1967 L total_room 1752 25187 4860 1474 2772 2098 1474 4907 1888 4967 1888 4967 1888 4967 1888 4967 1888 4967 1888 4967 1888 4967 1888 4967 1888 4967 1888 4967 1933 1955 1977 1977	
691 1 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275	984 A Sr. No. ⇒ 302 330 338 340 386 340 386 340 386 351 401 462 523 542 655 663 684 738 750 750	B households 〒 336 3478 1073 401 584 402 341 396 915 1158 1158 1158 1157 414 474 474 474	C housing_media 7 19 5 2 2 6 17 16 15 20 9 14 31 16 37 30 0 22	Jatitude ▼ 38.11 34.26 38.81 39.13 39.13 39.13 39.13 39.13 33.7.31 38.25 38.46 41.32 39.44 38.49 34.99 41.3	-117.73 E longitude -122.6 -118.9 -123.2 -123.21 -120.44 -121.8 -122.82 -122.79 -122.73 -122.85 -123.79 -122.85 -123.79 -122.85 -123.79 -122.85 -123.85 -123.79 -122.81 -122.85 -123.79 -122.81 -123.85 -	118100 F 201600 321300 151900 84400 142100 164500 189100 188600 138600 38700 118200 1146200 1146500 038700 116200 158000 158000 62700	G median_incoms ¥ 3.8068 6.8712 4.5042 2.1509 4.5045 2.875 5.038 2.0651 1.0486 3.2171 3.343 5.0286 1.8065	INLAND H ocean_proximity <1H OCEAN	I OCEAN I OCEAN I OCEAN INLAND	1459 population - 873 11966 2332 1065 1501 1252 1277 826 22581 2323 425 1177 1137 999 999 686	x78 K total_bedroomt 328 3521 1193 417 576 558 378 4111 885 1298 423 1298 423 519 3424 372	1967 L total_rooms ₹ 1752 25187 4860 1474 2772 2098 1807 1807 1807 1807 1808 4967 1975 2466 1975 2466 1975 1580	
691 691 1 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275 280	984 A 302 330 338 340 366 371 401 462 553 642 655 6683 684 738 750 762	B bouseholds 〒 336 3478 1073 401 584 492 341 396 915 1158 1158 157 414 474 424 320 264 34	C housing.media = 19 5 2 2 2 2 2 2 2 3 4 19 10 10 10 10 10 10 10 10 10 10 10 10 10	Jatitude マ 38.11 34.26 33.65 39.13 39.18 34.93 37.31 39.18 38.42 38.46 38.42 38.42 38.42 38.42 38.43 34.43 38.44 38.49 34.43 34.49 34.43 34.43 34.43 32.83	-117.73 E longitude -122.6 -118.9 -123.21 -123.21 -123.44 -121.8 -122.62 -122.79 -122.79 -122.79 -122.79 -122.79 -122.85 -123.85 -123.85 -123.79 -122.85 -123.85 -123.66 -126.76 -126.76	118100 F median. house. ₹ 201600 321300 151900 84400 142100 164500 188100 188500 138400 138400 118200 146500 146500 146500 146500 148500 1	G median_incoms ¥ 3.8068 6.8712 4.5022 1.375 2.6275 2.1509 4.5045 2.875 5.038 2.0651 1.0486 3.2171 3.3343 5.0286 1.8065 2.6458	INLAND H ocean, proximity <1H OCEAN	<th i="" i<="" inland="" ocean="" td=""><td>J population 〒 873 11956 2332 1065 1501 1252 1277 826 2281 2281 1177 1137 999 666 101 666 101 101</td><td>x78 K total_bedroomt = 328 3521 1193 378 378 378 417 885 1298 423 519 342 372 372 372</td><td>1967 Lotal_room 〒 1752 25187 4860 1474 2772 2099 1808 4967 1808 4967 1808 4967 1808 4967 1808 4967 1808 4967 1958 4967 1957 19</td></th>	<td>J population 〒 873 11956 2332 1065 1501 1252 1277 826 2281 2281 1177 1137 999 666 101 666 101 101</td> <td>x78 K total_bedroomt = 328 3521 1193 378 378 378 417 885 1298 423 519 342 372 372 372</td> <td>1967 Lotal_room 〒 1752 25187 4860 1474 2772 2099 1808 4967 1808 4967 1808 4967 1808 4967 1808 4967 1808 4967 1958 4967 1957 19</td>	J population 〒 873 11956 2332 1065 1501 1252 1277 826 2281 2281 1177 1137 999 666 101 666 101 101	x78 K total_bedroomt = 328 3521 1193 378 378 378 417 885 1298 423 519 342 372 372 372	1967 Lotal_room 〒 1752 25187 4860 1474 2772 2099 1808 4967 1808 4967 1808 4967 1808 4967 1808 4967 1808 4967 1958 4967 1957 19
691 691 1 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275 280 289	A A Sr. No. 302 330 338 340 366 371 401 462 523 542 6653 6683 738 750 762 778	B households 336 3478 1073 401 584 492 396 915 1158 1158 157 414 474 474 320 284 341 158 157 414 474 474 474 474 474 474 47	C housing_media ∓ 19 5 2 2 2 2 2 6 17 7 16 5 2 0 9 9 114 31 14 31 37 30 22 23 3 37	Jatitude ▼ 1atitude ₹ 38.11 34.68 33.65 39.13 39.13 39.13 34.03 37.31 38.42 38.42 38.42 38.42 38.46 41.32 39.44 38.49 34.49 34.49 34.49 34.89 33.22.83 32.283	-117.73 E longitude -122.6 -118.9 -123.21 -123.21 -123.21 -122.42 -122.42 -122.73 -122.73 -122.73 -122.82 -122.73 -122.83 -122.84 -123.86 -123.95	118100 F median. house. ₹ 201600 321300 151900 84400 142100 142100 148500 185600 135400 115200 1	G median_incomi ¥ 3.8068 6.9712 4.5022 1.375 2.875 2.875 2.875 5.038 2.0651 1.0488 3.2171 3.8343 5.0286 1.8065 2.4458 2.9643	INLAND H ocean_proximity <1H OCEAN	<th ocean<="" td=""><td>1459 population </td><td>378 k total_bedroomi = 328 3521 1193 417 576 358 378 411 855 1288 423 519 342 372 228 2292 221</td><td>ل total_rooms 25183 4860 1474 2775 2096 1807 1888 4967 4968 2017 2466 938 2017 2466 1975 1586 1454 1044 1044 1044 3771</td></th>	<td>1459 population </td> <td>378 k total_bedroomi = 328 3521 1193 417 576 358 378 411 855 1288 423 519 342 372 228 2292 221</td> <td>ل total_rooms 25183 4860 1474 2775 2096 1807 1888 4967 4968 2017 2466 938 2017 2466 1975 1586 1454 1044 1044 1044 3771</td>	1459 population	378 k total_bedroomi = 328 3521 1193 417 576 358 378 411 855 1288 423 519 342 372 228 2292 221	ل total_rooms 25183 4860 1474 2775 2096 1807 1888 4967 4968 2017 2466 938 2017 2466 1975 1586 1454 1044 1044 1044 3771
691 1 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275 280 289 311	A Sr. No. ⇒ 302 330 338 340 396 371 401 462 683 683 684 738 738 750 778 816	B households 〒 336 3478 1073 401 584 492 396 915 1158 1158 1157 414 474 474 320 264 348 196 320 264 348 348 348 348 320 320 320 320 320 320 320 320	C housing_media ╤ 19 5 2 2 2 2 2 2 2 3 1 4 3 3 1 6 3 3 1 6 3 3 0 2 2 2 2 3 3 7 3 3 2 2 2 3 3 3 2 3 3 3 3	Jatitude	-117.73 E longitude -122.6 -118.39 -123.22 -123.22 -123.22 -122.42 -122.42 -122.73 -122.43 -122.43 -122.43 -122.43 -122.43 -122.43 -122.43 -122.43 -128.45 -118.25 -118.25 -118.25 -128.45 -128.45 -128.45 -129.45 -118.25 -118.25 -118.25 -118.25 -129.45 -118.25 -118.25 -129.45 -129.45 -118.25 -118.25 -129.45 -129.45 -129.45 -118.25 -129.45 -129.45 -129.45 -129.45 -129.45 -118.25 -129.45 -129.45 -129.45 -129.45 -118.25 -129.45	118100 F median, house, ₹ 201600 321300 84400 151900 84400 142100 164500 189100 189100 135400 135400 118200 118200 1158000 112500 1	G median_incoms ¥ 3.8068 6.9712 4.5022 1.375 2.8275 2.8275 2.875 2.875 3.0851 1.0488 3.20851 1.0488 3.2171 3.6343 5.0286 1.8065 1.8065 2.8458 3.2415	INLAND H ocean_proximity <1H OCEAN	<th i="" i<="" inland="" ocean="" td=""><td>1459 population 〒 873 11956 2332 1355 1501 1252 1277 826 2383 425 1177 1137 999 666 1001 099 1786</td><td>378 K 1103 3521 1193 417 576 558 378 411 885 1288 423 519 342 372 322 323 423 519 342 372 32 213 741</td><td>1967 L total rooms 25187 4866 1474 2775 2098 1800 1900 1800 1900 1800 1900 1800 1900 1000</td></th>	<td>1459 population 〒 873 11956 2332 1355 1501 1252 1277 826 2383 425 1177 1137 999 666 1001 099 1786</td> <td>378 K 1103 3521 1193 417 576 558 378 411 885 1288 423 519 342 372 322 323 423 519 342 372 32 213 741</td> <td>1967 L total rooms 25187 4866 1474 2775 2098 1800 1900 1800 1900 1800 1900 1800 1900 1000</td>	1459 population 〒 873 11956 2332 1355 1501 1252 1277 826 2383 425 1177 1137 999 666 1001 099 1786	378 K 1103 3521 1193 417 576 558 378 411 885 1288 423 519 342 372 322 323 423 519 342 372 32 213 741	1967 L total rooms 25187 4866 1474 2775 2098 1800 1900 1800 1900 1800 1900 1800 1900 1000
691 1 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275 280 289 311 335	A Sr. No. ⇒ 302 330 338 340 398 340 398 340 398 350 398 361 401 462 523 542 655 683 684 738 750 752 778 816 847 847	B households	C housing_media ⊽ 19 5 2 2 6 17 16 15 20 9 14 3 1 1 16 3 3 7 30 0 22 2 23 3 3 7 32 2 16	Jatitude ▼ 38.11 34.63 33.65 33.65 33.913 39.13 39.13 37.31 38.46 41.32 39.44 38.46 41.32 39.44 38.49 34.89 41.3 32.83 33.84 38.44	-117.73	118100 F 201600 321300 151900 84400 142100 164500 189100 185600 189100 135400 135400 135400 135400 135600 135700 135600 135700 135600 135700 135700 135600 1357000 1357000 1357000 1357000 1357000 1357000 1357000 1357000 1357000 13570000 13570000 13570000 1357000000000000000000000000000000000000	G median_incomx ¥ 3.8068 6.8712 4.5042 2.1509 4.5045 2.875 5.038 2.0651 1.0486 3.2171 3.3434 5.0286 1.8065 2.4458 2.4458 2.86458	INLAND H ocean_proximity <1H OCEAN	<th ocean<="" td=""><td>1459 population 〒 873 11966 2332 1255 1501 1252 1277 826 22581 2323 425 1177 1137 999 686 101 101 109 109 109 109 109 109</td><td>378 K total_bedroomt = 328 3521 1193 417 576 558 378 411 885 1298 423 519 342 372 322 423 711 608</td><td>1967 L total_rooms 25187 48600 1474 2772 2098 1807 1807 1888 4967 1888 4967 2017 2469 1979 1958 1979 1958 1979 2453 1974 1987 1977 1988 1977 1988 1977 1977 1988 1977 1977 1988 1977 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977</td></th>	<td>1459 population 〒 873 11966 2332 1255 1501 1252 1277 826 22581 2323 425 1177 1137 999 686 101 101 109 109 109 109 109 109</td> <td>378 K total_bedroomt = 328 3521 1193 417 576 558 378 411 885 1298 423 519 342 372 322 423 711 608</td> <td>1967 L total_rooms 25187 48600 1474 2772 2098 1807 1807 1888 4967 1888 4967 2017 2469 1979 1958 1979 1958 1979 2453 1974 1987 1977 1988 1977 1988 1977 1977 1988 1977 1977 1988 1977 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977</td>	1459 population 〒 873 11966 2332 1255 1501 1252 1277 826 22581 2323 425 1177 1137 999 686 101 101 109 109 109 109 109 109	378 K total_bedroomt = 328 3521 1193 417 576 558 378 411 885 1298 423 519 342 372 322 423 711 608	1967 L total_rooms 25187 48600 1474 2772 2098 1807 1807 1888 4967 1888 4967 2017 2469 1979 1958 1979 1958 1979 2453 1974 1987 1977 1988 1977 1988 1977 1977 1988 1977 1977 1988 1977 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977 1978 1977
691 691 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275 280 289 311 335 370	A A 302 330 338 340 366 371 401 462 523 542 665 6683 750 778 778 816 647 924	B bouseholds ₹ 336 3478 1073 401 584 492 341 396 915 1158 1157 414 474 320 224 341 196 721 586 712	C housing.media ╤ 19 5 2 2 2 2 2 2 3 19 19 10 10 10 10 10 10 10 10 10 10 10 10 10	Jatitude ▼ 38.01 34.03 34.03 34.03 33.65 39.13 39.18 34.93 37.31 39.18 34.93 37.31 38.42 38.42 38.44 38.49 34.43 34.49 34.43 34.49 34.43 38.49 34.3 32.83 33.83 33.89 38.844 38.46 33.83 38.44	-117.73 E longitude -122.6 -118.9 -123.21 -123.21 -123.44 -121.8 -122.62 -122.79 -122.79 -122.86 -122.66 -116.97 -122.86 -116.97 -122.86 -112.67 -122.86 -112.67 -122.86 -112.87 -122.86 -122.87 -122.86 -122.86 -122.86 -112.87 -122.86 -116.97 -122.85 -122.85 -122.86 -116.97 -122.85	118100 F median. house. ₹ 201600 321300 151900 84400 142100 164500 188500 138400 138400 118200 118200 118200 146500 118200 118200 112500 1	G median_incoms ¥ 3.8068 6.8712 4.5022 1.375 2.6275 2.1509 4.5045 2.875 5.038 2.0651 1.0488 3.2171 3.38343 5.0286 1.9068 2.8458 2.9843 3.2415 2.6354	INLAND H ocean_proximity <1H OCEAN	<th i="" inl<="" inland="" ocean="" td=""><td>J population ₹ 873 11956 2332 1065 1501 1252 1252 1267 2333 425 1177 1137 1137 1137 999 668 101 699 1766 1683 1970 1970</td><td>X78 K total. bedroomt = 328 3521 1193 417 576 558 378 411 885 1288 423 519 342 372 32 213 741 606 817</td><td>1967 L total_room ₹ 25187 25187 25187 2008 1427 2008 1497 1688 4967 4062 938 2017 2469 1979 1580 1494 1042 3771 1589 1494 1042 3771 1589 1495 14</td></th>	<td>J population ₹ 873 11956 2332 1065 1501 1252 1252 1267 2333 425 1177 1137 1137 1137 999 668 101 699 1766 1683 1970 1970</td> <td>X78 K total. bedroomt = 328 3521 1193 417 576 558 378 411 885 1288 423 519 342 372 32 213 741 606 817</td> <td>1967 L total_room ₹ 25187 25187 25187 2008 1427 2008 1497 1688 4967 4062 938 2017 2469 1979 1580 1494 1042 3771 1589 1494 1042 3771 1589 1495 14</td>	J population ₹ 873 11956 2332 1065 1501 1252 1252 1267 2333 425 1177 1137 1137 1137 999 668 101 699 1766 1683 1970 1970	X78 K total. bedroomt = 328 3521 1193 417 576 558 378 411 885 1288 423 519 342 372 32 213 741 606 817	1967 L total_room ₹ 25187 25187 25187 2008 1427 2008 1497 1688 4967 4062 938 2017 2469 1979 1580 1494 1042 3771 1589 1494 1042 3771 1589 1495 14
691 691 3 18 25 27 41 46 68 105 144 152 218 233 234 267 275 280 289 311 335 370 387	A A Sr. No. 330 338 338 338 338 338 338 338 338 338 366 371 462 523 665 665 665 665 665 738 778 616 847 952	B households 〒 336 3478 1073 401 584 402 396 915 1158 1158 1158 1157 414 474 474 474 320 224 349 224 349 157 157 414 757 414 757 414 757 414 757 757 757 757 757 757 757 75	C housing_media ╤ 19 5 2 2 2 2 2 2 3 3 1 4 3 1 4 3 1 1 6 3 7 3 2 2 2 3 3 3 7 3 2 2 2 3 3 7 3 2 2 2 3 3 7 3 2 2 2 3 3 7 3 2 2 3 3 3 7 3 2 3 3 3 7 3 2 3 3 3 3	Jatitude	-117.73 E longitude -122.6 -118.9 -117.59 -123.2 -123.2 -123.2 -123.2 -122.42 -122.43 -122.73 -122.73 -122.85 -123.79 -122.43 -122.43 -122.43 -122.43 -122.43 -128.45 -118.25 -122.67 -118.25 -122.67 -122.75 -118.25 -118.24 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -122.45 -128.45 -12	118100 F median, house, ⇒ 201600 321300 84400 142100 144500 144500 148500 138400 138400 138400 138400 148500 142500 140500 1	G median_Incomx ¥ 3.8068 6.9712 4.5022 1.375 2.2275 2.1509 4.5045 2.875 5.038 2.0651 1.0486 3.32171 3.6343 5.0286 1.8055 2.6453 3.32415 2.3344 6.1078 2.4107	INLAND H ocean_proximity <1H OCEAN	<th i="" inl<="" inland="" ocean="" td=""><td>1459 population → 873 873 11956 2332 1065 1501 1252 1277 826 2281 1277 826 2283 425 1177 1137 999 666 1001 1078 699 1786 699 1786 699 1786 699 1786 699 1970 498</td><td>378 K 1103 3521 1193 417 576 558 378 411 885 1298 423 519 342 372 322 213 741 606 817</td><td>1967 L total_rooms 〒 25187 25187 25187 1752 2098 1474 2772 2098 1807 1888 4967 2469 2017 2469 1979 1580 1497 1580 1492 2469 1979 1580 1492 2469 1979 1580 1494 2618 1474 1975 1975 1987 1979 1977 19</td></th>	<td>1459 population → 873 873 11956 2332 1065 1501 1252 1277 826 2281 1277 826 2283 425 1177 1137 999 666 1001 1078 699 1786 699 1786 699 1786 699 1786 699 1970 498</td> <td>378 K 1103 3521 1193 417 576 558 378 411 885 1298 423 519 342 372 322 213 741 606 817</td> <td>1967 L total_rooms 〒 25187 25187 25187 1752 2098 1474 2772 2098 1807 1888 4967 2469 2017 2469 1979 1580 1497 1580 1492 2469 1979 1580 1492 2469 1979 1580 1494 2618 1474 1975 1975 1987 1979 1977 19</td>	1459 population → 873 873 11956 2332 1065 1501 1252 1277 826 2281 1277 826 2283 425 1177 1137 999 666 1001 1078 699 1786 699 1786 699 1786 699 1786 699 1970 498	378 K 1103 3521 1193 417 576 558 378 411 885 1298 423 519 342 372 322 213 741 606 817	1967 L total_rooms 〒 25187 25187 25187 1752 2098 1474 2772 2098 1807 1888 4967 2469 2017 2469 1979 1580 1497 1580 1492 2469 1979 1580 1492 2469 1979 1580 1494 2618 1474 1975 1975 1987 1979 1977 19

Images of Sheets : Screenshot of Prediction anomalies noticed in the dataset h_3000.csv

Description: These screenshots depict the incorrectly classified. The ones which were actually INLAND but were predicted as <1H OCEAN and the one which were <1H OCEAN and predicted as INLAND.

Qualitative Description of Error Analysis on Housing Ocean Proximity Data

The error analysis of the LightSide logistic regression model, tasked with predicting 'ocean_proximity', reveals discernible patterns in its predictive performance. The confusion matrices across the screenshots show varying degrees of misclassification between the "<1H OCEAN" and "INLAND" categories, particularly noted by the number of instances classified as "<1H OCEAN" when they are actually "INLAND", and vice versa. For example, in one confusion matrix, I observed nearly 160 true positives for "<1H OCEAN" but also 21 false negatives, where "<1H OCEAN" instances are mislabeled as "INLAND". Conversely, there are 107 true positives for "INLAND" but 12 false negatives, indicating a misclassification of "INLAND" instances as "<1H OCEAN".

The feature importance table offers additional insights, showcasing the strong influence of features such as 'median_income', 'latitude', and 'longitude' on the predictions, which is to be expected given their direct relevance to the concept of ocean proximity. However, the model's kappa score, which is a more robust measure than accuracy since it accounts for random chance, varies slightly but hovers around 0.773 in one instance, suggesting moderate agreement. An accuracy of 0.89 implies that the model correctly predicts 89% of the instances, but the kappa score indicates that the model's predictive power is less than perfect when adjusted for chance.

The average cell values in the confusion matrices, like the 253710.7 for "<1H OCEAN" and 142700 for "INLAND", along with the significant horizontal absolute differences (e.g., 121464.842 for "<1H OCEAN"), highlight the disparity in prediction errors between classes. These values suggest that for some instances, the model's confidence in its predictions is not consistent across the board.

Potential Solutions

In conclusion, the model demonstrates commendable performance; however, error analysis reveals a distinct tendency towards specific misclassifications. These could potentially be mitigated by rectifying inherent biases favoring over-represented classes or by enhancing the decision boundaries delineation among classes. Augmenting the model with a dataset that ensures class balance, coupled with the incorporation of additional features encapsulating geographic distribution nuances, is anticipated to diminish the observed predictive inaccuracies and consequently elevate the kappa statistic.

Tuning

Tuning methodology in Weka involves adjusting various parameters of machine learning algorithms to optimize their performance and for this process is critical in achieving more accurate and efficient models. Weka provides tools and interfaces for tuning, such as Explorer and Experimenter, allowing users to experiment with different parameter settings. The objective is to find the best combination of parameters that yield the highest accuracy or other performance metrics on given datasets.

Preprocess Classify Cluster Ass	sociate Select attributes Visualize
Classifier	
Choose SMO -C 1.0 -L 0.001 -P 1.0E-	-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functi
Test options	Classifier output
O Use training set	+ -1.4328 * (normalized) population + -0.5887 * (normalized) households
Supplied test set Set	+ 1.1885 * (normalized) median_income
Cross-validation Folds 10	+ -4.9497 * (normalized) median_house_value - 2.9463
O Percentage split % 66	
More options	Number of kernel evaluations: 23809 (69.374% cached)
(Nom) ocean_proximity ~	Time taken to build model: 0.03 seconds
Start Stop	=== Stratified cross-validation ===
Result list (right-click for options)	=== Summary ===
20:47:37 - functions.SMO	Correctly Classified Instances 644 92 %
	Incorrectly Classified Instances 56 8 % Kappa statistic 0.8334 Mean absolute error 0.08 Root mean squared error 0.2828 Relative absolute error 16.4342 % Root relative squared error 57.3328 % Total Number of Instances 700
	=== Detailed Accuracy By Class ===
	TP Rate FP Rate Precision Recall F-Measure MCC ROC Area Class 0.966 0.143 0.903 0.966 0.933 0.836 0.911 0.892 <1H OCEAN 0.857 0.034 0.947 0.857 0.900 0.836 0.911 0.871 INLAND Weighted Avg. 0.920 0.922 0.920 0.919 0.836 0.911 0.884 === Confusion Matrix === <
	a b < classified as 393 14 a = <1H OCEAN 42 251 b = INLAND

Initial Sets Baseline Performance

A Sequential Minimal Optimization (SMO) classifier was used to build a support vector machine model with a polynomial kernel. The model was validated using 10-fold cross-validation on a dataset of 700 instances. The classifier achieved a high classification accuracy, correctly predicting 92% of the instances. The Kappa statistic of 0.8334 suggests a strong agreement beyond chance. The detailed accuracy by class shows a true positive rate of 0.966 for the "<1H OCEAN" class and 0.857 for the "INLAND" class, indicating a slightly better performance for the former. The confusion matrix shows that the model had more difficulty distinguishing the "INLAND" class, with 42 instances of "<1H OCEAN" being misclassified as "INLAND". Overall, the weighted average F-Measure of 0.919 and ROC area of 0.911 reflect a robust model performance.

Classifier Choose SMO -C 1.0 -L 0.001 -P 1.0	
Chasse SMO -C 1 0 -L 0 001 -P 1 0	
CHOOSE SINC -C 1.0 -L 0.001 -F 1.0	E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.funct
Test options Use training set Supplied test set Cross-validation Folds 10	Classifier output + -1.3877 * (normalized) population + -0.6767 * (normalized) households + 0.9054 * (normalized) median_income + -4.6852 * (normalized) median_house_value
Percentage split % 66	 – 2.8272 Number of kernel evaluations: 22354 (72.311% cached)
More options	
(Nom) ocean_proximity	Time taken to build model: 0.01 seconds
Start Stop	=== Stratified cross-validation ===
Result list (right-click for options)	=== Summary ===
20:47:37 - functions.SMO 20:50:54 - functions.SMO	Correctly Classified Instances 580 92.0635 % Incorrectly Classified Instances 50 7.9365 % Kappa statistic 0.8349 Mean absolute error 0.0794 Root mean squared error 0.2817 Relative absolute error 16.2983 % Root relative squared error 57.0953 % Total Number of Instances 630 === Detailed Accuracy By Class ===
	TP Rate PP recision Recall F-Measure MCC ROC Area PRC Area Class 0.964 0.140 0.905 0.964 0.934 0.838 0.912 0.894 <1H OCEAN 0.860 0.036 0.946 0.860 0.901 0.838 0.912 0.894 <1H OCEAN Weighted Avg. 0.921 0.0966 0.922 0.921 0.920 0.838 0.912 0.885 === Confusion Matrix === a b < classified as 353 13 a = <1H OCEAN 37 227 b = INLAND

Performing Tuning	Analysis: On four Training and Testing Set using StratefiedRemoveFolds filter

Train Set 1

Description: The SMO (Sequential Minimal Optimization) algorithm was employed to train a support vector machine with a polynomial kernel. The model was tested using 10-fold cross-validation on a dataset comprising 630 instances. It achieved a commendable classification accuracy, correctly classifying 92.0635% of instances and incorrectly classifying 7.9365%. The Kappa statistic of 0.8349 signifies a substantial agreement beyond chance. When observing the detailed accuracy by class, the true positive rate (TP Rate) for the "<1H OCEAN" class was 0.964, with a precision of 0.905, and for the "INLAND" class, the TP Rate was 0.860, with a precision of 0.946. The confusion matrix indicates that the model had a higher tendency to misclassify the "INLAND" class as "<1H OCEAN" with 37 instances misclassified. The F-Measure of 0.920 and ROC area of 0.912 further indicate a strong predictive performance of the model. The number of kernel evaluations was 22354, with 72.311% cached, **suggesting an efficient computational process.**

Preprocess Classify Cluster Ass	ociate Select attributes Visualize
Classifier Choose SMO -C 1.0 -L 0.001 -P 1.0E-	12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions
Test options Use training set Supplied test set Cross-validation Folds 10 Percentage split % 66 More options	Classifier output > 6 (numeric) population (numeric) population > 6 (numeric) population (numeric) households > 7 (numeric) households (numeric) median_income > 8 (numeric) median_income (numeric) median_house_value > 9 (numeric) median_house_value (nominal) ocean_proximity > 10 (nominal) ocean_proximity
(Nom) ocean_proximity Start Stop Result list (right-click for options) 20:47:37 - functions.SMO 20:50:54 - functions.SMO 20:52:34 - misc.InputMappedClassifier	<pre>=== Evaluation on test set === Time taken to test model on supplied test set: 0.03 seconds === Summary === Correctly Classified Instances 59 93.6508 % Incorrectly Classified Instances 4 6.3492 % Kappa statistic 0.866 Mean absolute error 0.0635 Root mean squared error 0.252 Relative absolute error 13.0666 % Root relative squared error 51.1769 % Total Number of Instances 63 === Detailed Accuracy By Class ===</pre>
	TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 1.000 0.154 0.902 1.000 0.949 0.874 0.923 0.902 1H OCEAN 0.846 0.000 1.000 0.846 0.917 0.874 0.923 0.902 1H OCEAN weighted Avg. 0.937 0.090 0.943 0.937 0.935 0.874 0.923 0.905 === Confusion Matrix === a b < classified as 37 0 a = <1H OCEAN 4 22 b = INLAND

Test Set 1

Description: The SMO (Sequential Minimal Optimization) algorithm was used to create a support vector machine model with a polynomial kernel. The model was evaluated on a test set of 63 instances, resulting in 93.6508% correctly classified instances and 6.3492% incorrectly classified instances. The Kappa statistic was 0.866, indicating a high level of agreement between the predicted and observed classifications. The mean absolute error was low at 0.0635, and the root mean squared error was 0.252, which are both indicators of the model's predictive accuracy.

The detailed accuracy by class shows that the model perfectly classified the "<1H OCEAN" class with a true positive rate of 1.000 and precision of 0.902. The "INLAND" class had a true positive rate of 0.846 and precision of 1.000, reflecting high accuracy but slightly less than the "<1H OCEAN" class. The weighted average for precision, recall, and F-Measure across classes was 0.943, 0.937, and 0.935, respectively, demonstrating **overall strong performance**.

The confusion matrix further reveals the model's performance, with no misclassifications for the "<1H OCEAN" class and only 4 instances of the "INLAND" class being misclassified as "<1H OCEAN". The model's robustness is also reflected in the high Matthews correlation coefficient (MCC) of 0.874 and the receiver operating characteristic (ROC) area of 0.923, which suggests a good balance between sensitivity and specificity. The relative absolute error and root relative squared error are relatively high at 13.0666% and 51.1769%, indicating areas where the model's performance could potentially be improved.

····, ·····	sociate Select attributes Visualize
Classifier Choose SMO -C 1.0 -L 0.001 -P 1.0E	-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.fur
Test options Use training set Supplied test set Cross-validation Percentage split More options (Nom) ocean_proximity Start Stop Result list (right-click for options) 20:47:37 - functions.SMO 20:50:54 - functions.SMO 20:52:34 - misc.InputMappedClassifier 20:53:29 - functions.SMO	Classifier output + -1.3028 * (normalized) population + -0.8136 * (normalized) households + 0.9395 * (normalized) median_income + -4.7665 * (normalized) median_house_value - 2.4749 Number of kernel evaluations: 19719 (67.937% cached) Time taken to build model: 0.01 seconds === Stratified cross-validation === === Summary === Correctly Classified Instances 521 91.8871 % Incorrectly Classified Instances 46 8.1129 % Kappa statistic 0.8322 Mean absolute error 0.2848 Relative absolute error 16.5793 % Root relative squared error 57.5853 % Total Number of Instances 567 === Detailed Accuracy By Class ===
	TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0.963 0.140 0.902 0.963 0.932 0.835 0.911 0.890 <1H OCE/ 0.860 0.037 0.945 0.860 0.900 0.835 0.911 0.873 INLAND Weighted Avg. 0.919 0.096 0.921 0.919 0.918 0.835 0.911 0.882 === Confusion Matrix === a b < classified as 313 12 a = <1H OCEAN 34 208 b = INLAND

TrainSet 2

Description: The SMO classifier with a polynomial kernel was applied to a dataset, evaluated using 10-fold stratified cross-validation. The model has performed well, correctly classifying 91.8871% of the 567 instances. The Kappa statistic is 0.8322, which is a very good score indicating a high degree of agreement.

For detailed class accuracy, the classifier achieved a true positive rate (TP Rate) of 0.963 for the "<1H OCEAN" class and 0.860 for "INLAND". Precision for the "<1H OCEAN" is slightly better than for "INLAND" (0.902 vs. 0.945), as is the F-Measure (0.932 vs. 0.900). The Matthews Correlation Coefficient (MCC) of 0.835 for both classes suggests a high-quality classifier.

The ROC Area under the curve is 0.911 for both classes, indicating a high true positive rate relative to the false positive rate. The PRC Area, which is the area under the precision-recall curve, is 0.890 for "<1H OCEAN" and 0.873 for "INLAND", which is also indicative of a **good predictive performance**.

The confusion matrix shows that the classifier has some difficulty distinguishing between the two classes, with 24 instances of "<1H OCEAN" being incorrectly classified as "INLAND" and 12 instances of "INLAND" being incorrectly classified as "<1H OCEAN".

The model seems to be effective and efficient, taking only 0.01 seconds to build, with a substantial portion of the kernel evaluations (67.937%) being cached, which helps in speeding up the computations. The errors such as mean absolute error (0.0811) and root mean squared error (0.2848) are relatively low, which complements the **high classification accuracy**.

	E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.fu
Test options Use training set Supplied test set Cross-validation Percentage split More options	Classifier output (numeric) population> 6 (numeric) population (numeric) households> 7 (numeric) households (numeric) median_income> 8 (numeric) median_income (numeric) median_house_value> 9 (numeric) median_house_value (nominal) ocean_proximity> 10 (nominal) ocean_proximity Time taken to build model: 0 seconds
	<pre>/ === Evaluation on test set ===</pre>
(Nom) ocean_proximity Start Stop Result list (right-click for options)	Time taken to test model on supplied test set: 0.08 seconds === Summary ===
20:47:37 - functions.SMO 20:50:54 - functions.SMO 20:52:34 - misc.InputMappedClassifier 20:53:29 - functions.SMO 20:54:58 - misc.InputMappedClassifier	Correctly Classified Instances46991.9608 %Incorrectly Classified Instances418.0392 %Kappa statistic0.8354Mean absolute error0.0804Root mean squared error0.2835Relative absolute error16.3497 %Root relative squared error57.0475 %Total Number of Instances510=== Detailed Accuracy By Class ===
	TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0.968 0.142 0.896 0.968 0.931 0.839 0.913 0.885 <1H OCEA 0.858 0.032 0.956 0.858 0.904 0.839 0.913 0.883 INLAND Weighted Avg. 0.920 0.092 0.920 0.919 0.839 0.913 0.884 === Confusion Matrix === a b < classified as 275 9 a = 1H OCEAN 32 194 b INLAND 5 19 10 10

Test Set 2

Description: An SMO classifier using a polynomial kernel. The classifier was tested on a set of 510 instances and it managed to correctly classify 91.9608%

of them. It incorrectly classified 8.0392% of the instances. The Kappa statistic is 0.8354, which indicates a strong agreement between the classifier's predictions and the actual labels.

The detailed accuracy by class shows a true positive rate (TP Rate) for the "<1H OCEAN" class of 0.968 and for the "INLAND" class of 0.858. The precision is higher for the "INLAND" class at 0.956 compared to 0.896 for "<1H OCEAN". The F-Measure, which is a balance between precision and recall, is 0.931 for "<1H OCEAN" and 0.904 for "INLAND". The Matthews Correlation Coefficient (MCC) is 0.839 for both classes, suggesting that the classifier's predictions are of high quality.

The ROC Area, representing the trade-off between the true positive rate and the false positive rate, is 0.913 for both classes, which is considered excellent. The PRC Area, or the area under the precision-recall curve, is 0.885 for "<1H OCEAN" and slightly lower for "INLAND" at 0.883.

The confusion matrix provides insight into classification errors; the classifier confused "<1H OCEAN" with "INLAND" 9 times, and "INLAND" with "<1H OCEAN" 32 times.

The model shows high effectiveness in classifying instances with a relatively balanced performance across both classes. The relative absolute error at 16.3497% and the root relative squared error at 57.0475% are aspects that could potentially be improved, but they do not overly detract from the strong performance indicated by the other metrics.

Choose SMO -C 1.0 -L 0.001 -P 1.0E	-12 -N 0 -V -1 -W 1 -I	K "weka.cla	ssifiers.fur	ctions.suppo	rtVector.P	olyKernel -E 1	.0 -C 250	007" -calibrat	or "weka.cla	assifiers.funct
Test options Use training set Supplied test set Set Cross-validation Folds 10 Percentage split % 66 More options	+ -0.7592 + 0.8342	* (normal	ized) hou ized) med ized) med	seholds ian_income ian_house_v						
(Nom) ocean_proximity ~ Start Stop Result list (right-click for options)	Time taken to b === Stratified === Summary ===	cross-vali								
20:47:37 - functions.SMO 20:50:54 - functions.SMO 20:52:34 - misc.lnputMappedClassifier 20:53:29 - functions.SMO 20:54:58 - misc.lnputMappedClassifier 20:55:50 - functions.SMO 20:56:54 - functions.SMO 20:56:54 - functions.SMO 20:57:16 - misc.lnputMappedClassifier	Correctly Class Incorrectly Cla Kappa statistic Mean absolute e Root mean squar Relative absolu Root relative s Total Number of === Detailed Ac	ssified Ir rror ed error te error quared err Instances	or	467 43 0.82 0.08 0.29 17.08 58.45 510	43 04 2 %	91.5686 8.4314				
	Weighted Avg. === Confusion M a b < 273 11 a 3	TP Rate 0.961 0.858 0.916	FP Rate 0.142 0.039 0.096	Precision 0.895 0.946 0.918	Recall 0.961 0.858 0.916	F-Measure 0.927 0.900 0.915	MCC 0.830 0.830 0.830	ROC Area 0.910 0.910 0.910	PRC Area 0.882 0.875 0.879	Class <1H OCEAN INLAND

Train Set 3

Description: The SMO classifier with a polynomial kernel function was executed on a dataset, with the evaluation performed through 10-fold cross-validation. The model correctly classified 91.5686% of the instances (467 out of 510), misclassifying 8.4314% (43 instances). The Kappa statistic is 0.8275, indicating a very good agreement between the classifier predictions and the actual data.

The detailed accuracy by class shows that the classifier has a true positive rate (TP Rate) for the "<1H OCEAN" class of 0.961 and for the "INLAND" class of 0.858. Precision for "<1H OCEAN" is slightly lower at 0.895 compared to 0.946 for "INLAND". The F-Measure is 0.927 for "<1H OCEAN" and 0.900 for "INLAND". The Matthews Correlation Coefficient (MCC) is 0.830, which is a high value indicating a strong correlation between observed and predicted classifications.

The ROC Area is 0.910 for both classes, which is considered to be excellent, reflecting a model that provides a good trade-off between true positive and false positive rates. The PRC Area, which is the precision-recall curve area, is 0.882 for "<1H OCEAN" and 0.875 for "INLAND", both of which are indicative of a strong performance.

The confusion matrix shows that the classifier has more difficulty distinguishing the "INLAND" class with 32 instances of "INLAND" being incorrectly classified as "<1H OCEAN" and only 11 instances of "<1H OCEAN" being incorrectly classified as "INLAND".

The results indicate **a highly effective model**, especially considering that the model was built in 0.01 seconds and the time taken to test on the supplied test set was only 0.08 seconds. The errors such as mean absolute error (0.0843) and root mean squared error (0.2904) are relatively low, which, along with the relative absolute error (17.082%) and root relative squared error (58.4503%), suggests that there might be room for further optimization but the current performance is already strong.

Classifier	sociate Select attributes Visualize -12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.func
Test options Use training set Supplied test set Cross-validation Percentage split More options	Classifier output (numeric) population (numeric) households (numeric) median_income (numeric) median_income (numeric) median_house_value (nominal) ocean_proximity Time taken to build model: 0.01 seconds
(Nom) ocean_proximity Start Stop Result list (right-click for options) 20:47:37 - functions.SMO 20:50:54 - functions.SMO 20:52:34 - misc.InputMappedClassifier 20:53:50 - functions.SMO 20:55:50 - functions.SMO 20:55:51 - functions.SMO 20:56:54 - functions.SMO 20:57:16 - misc.InputMappedClassifier	<pre>=== Evaluation on test set === Time taken to test model on supplied test set: 0.01 seconds === Summary === Correctly Classified Instances 47 92.1569 % Incorrectly Classified Instances 4 7.8431 % Kappa statistic 0.8404 Mean absolute error 0.0784 Root mean squared error 0.2801 Relative absolute error 15.8624 % Root relative squared error 55.275 % Total Number of Instances 51 === Detailed Accuracy By Class ===</pre>
	TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0.964 0.130 0.900 0.964 0.931 0.843 0.917 0.887 <1H OCEAN 0.870 0.036 0.952 0.870 0.909 0.843 0.917 0.887 INLAND Weighted Avg. 0.922 0.088 0.924 0.922 0.921 0.843 0.917 0.887 == a b < classified as 27 1 a = <1H OCEAN 3 20 b = INLAND

TestSet 3

Description: The SMO classifier with a polynomial kernel has been applied to a dataset. The classifier was assessed on a test set consisting of 51 instances and achieved an accuracy of 92.1569%, correctly classifying 47 instances and incorrectly classifying 4. The Kappa statistic is 0.8404, suggesting a very good agreement.

The classifier's detailed accuracy by class shows that it had a true positive rate (TP Rate) of 0.964 for the "<1H OCEAN" class and 0.870 for the "INLAND" class. Precision is high for both classes, at 0.900 for "<1H OCEAN" and 0.952 for "INLAND". The F-Measure is also high, at 0.931 and 0.909 respectively, indicating a balanced harmonic mean of precision and recall. The Matthews Correlation Coefficient (MCC) stands at 0.843 for both classes, which is a high value indicating strong predictive performance.

The ROC Area, which measures the trade-off between true positive rate and false positive rate, is 0.917 for both classes, indicating a very good predictive ability. The PRC Area, representing the precision-recall curve, is also high, at 0.887 for "<1H OCEAN" and 0.887 for "INLAND".

The confusion matrix provides additional detail on the classification performance, revealing that the classifier did not misclassify any "<1H OCEAN" instances as "INLAND" (0 instances) but misclassified 3 "INLAND" instances as "<1H OCEAN". Therefore,

The model exhibits excellent performance with rapid processing times, as indicated by the 0.01 seconds taken to build the model and the same time to test it on the supplied test set. The mean absolute error is low at 0.0784, and the root mean squared error is 0.2801, further indicating the model's accuracy. However, the relative absolute error and root relative squared error are somewhat high at 15.8624% and 56.2757%, respectively, which could suggest areas for potential improvement in model calibration or feature selection.

Test options	Classifier output									
Use training set		* (normal	ized) pop	ulation						
-	+ -0.8513 * (normalized) households + 0.6329 * (normalized) median_income + -4.5379 * (normalized) median house value									
Supplied test set Set										
Cross-validation Folds 10	- 2.0633	* (normat	.izeu) met	itan_nouse_v	acue					
Percentage split % 66	Number of Issues		1007							
More options	Number of kernel	t evaluati	ons: 1083	67 (73,902%	cached)					
Nom) ocean_proximity	Time taken to bu	uild model	: 0 secor	ıds						
Start Stop	=== Stratified of	ross-vali	dation ==	=						
esult list (right-click for options)	=== Summary ===									
20:47:37 - functions.SMO	Correctly Classi	ified Inst	ances	421		91.7211	%			
20:50:54 - functions.SMO	Incorrectly Classified Instances 38 8.2789 %									
20:52:34 - misc.InputMappedClassifier	Kappa statistic 0.831									
20:53:29 - functions.SMO	Mean absolute error 0.0828 Root mean squared error 0.2877									
20:54:58 - misc.InputMappedClassifier	Relative absolute error 16.7316 %									
20:55:50 - functions.SMO	Root relative squared error 57.8475 %									
20:56:54 - functions.SMO	Total Number of	Instances		459						
20:57:16 - misc.InputMappedClassifier 20:58:55 - functions.SMO	=== Detailed Accuracy By Class ===									
20:59:08 - misc.InputMappedClassifier				Precision		F-Measure			PRC Area	Class
		0.968 0.854	0.146 0.032	0.891 0.957	0.968 0.854	0.928 0.903	0.835 0.835	0.911 0.911	0.880 0.883	<1H OCEAN INLAND
	Weighted Avg.	0.034	0.032	0.920	0.034	0.903	0.835	0.911	0.881	INLAND
	=== Confusion Ma	atrix ===								
	a b < 0	classified	as							
	245 8 a =	= <1H OCEA = INLAND	Ν							

Train Set 4

Description: The SMO classifier with a polynomial kernel shows that the model was trained on a dataset with 459 instances and evaluated using 10-fold stratified cross-validation. The classifier achieved an accuracy of 91.7211%, correctly classifying 421 instances while incorrectly classifying 38. The Kappa statistic is 0.831, indicating a strong level of agreement between the classifier's predictions and the actual class labels.

The detailed accuracy by class shows a high true positive rate (TP Rate) for the "<1H OCEAN" class at 0.968 and for the "INLAND" class at 0.854. The classifier demonstrated high precision, particularly for the "INLAND" class at 0.957, and a balanced F-Measure of 0.928 for "<1H OCEAN" and 0.903 for "INLAND". The Matthews Correlation Coefficient (MCC) of 0.835 for both classes indicates a high-quality prediction.

The area under the ROC curve (ROC Area) is 0.911 for both classes, suggesting excellent discriminatory ability. The area under the precision-recall curve (PRC

Area) is also high at 0.880 for "<1H OCEAN" and 0.883 for "INLAND", indicating a strong precision and recall balance.

The confusion matrix shows that the classifier predicted the "<1H OCEAN" class with few errors (8 instances misclassified as "INLAND"), and more instances of the "INLAND" class were misclassified as "<1H OCEAN" (30 instances). Therefore,

The performance metrics suggest that the **classifier is highly effective**, with a rapid model build time of 0 seconds. The mean absolute error is small at 0.0828, and the root mean squared error at 0.2877 is low, which are indicative of a model with accurate predictions. The relative absolute error and root relative squared error are moderately high at 16.7316% and 57.8475%, respectively, but these do not significantly detract from the **overall strong performance of the classifier.**

Preprocess Classify Cluster As	sociate Select attributes Visualize
	-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.fun
Test options Use training set Supplied test set Cross-validation Folds Percentage split % 66 More options (Nom) ocean_proximity Start Stop	Classifier output (numeric) population> 6 (numeric) population (numeric) households> 7 (numeric) households (numeric) median_income> 8 (numeric) median_income (numeric) median_house_value> 9 (numeric) median_house_value (nominal) ocean_proximity> 10 (nominal) ocean_proximity Time taken to build model: 0 seconds === Evaluation on test set === Time taken to test model on supplied test set: 0.01 seconds
Result list (right-click for options) 20:47:37 - functions.SMO 20:50:54 - functions.SMO 20:52:34 - misc.InputMappedClassifier 20:53:29 - functions.SMO 20:54:58 - misc.InputMappedClassifier 20:55:50 - functions.SMO 20:57:16 - misc.InputMappedClassifier 20:58:55 - functions.SMO 20:59:08 - misc.InputMappedClassifier	<pre>=== Summary === Correctly Classified Instances</pre>

Test Set 4

Description: An SMO (Sequential Minimal Optimization) model using a polynomial kernel was applied to a dataset, producing an accuracy of 91.3043% on a test set of 46 instances. The Kappa statistic stands at 0.8189, indicating a strong agreement between the predicted and actual classifications. The model shows perfect recall for the "<1H OCEAN" class, with a true positive rate (TP Rate) of 1.000, although with a false positive rate (FP Rate) of 0.200, suggesting some instances of other classes were incorrectly labeled as "<1H OCEAN". The precision for this class was 0.867, and the F-Measure, which combines precision and recall, was 0.929. The model also performed well for the "INLAND" class, with a TP Rate of 0.800 and a precision of 1.000, indicating no instances were wrongly labeled as "INLAND". The Matthews Correlation Coefficient (MCC) for both classes was 0.833, showing a high-guality prediction power. The model's evaluation on the test set was completed remarkably quickly, in just 0.01 seconds. Despite the high accuracy, there is room for improvement, particularly in reducing the false positive rate for the "<1H OCEAN" class, as reflected by the confusion matrix

where 4 "INLAND" instances were misclassified. The mean absolute error and root mean squared error are low at 0.087 and 0.2949, respectively, but the relative errors are moderately high, suggesting potential areas for model refinement.

Final Conclusion about performing Tuning:

Arguments for tuning:

While tuning consistently nudged accuracy upwards across all train sets, balanced performance across both classes, and slightly reduced misclassifications, the overall improvements were small (around 0.2% accuracy increase). This suggests that here tuning's value depends on the importance of slight accuracy gains and model robustness, and whether the potential benefits outweigh the additional time and resources required.

Arguments against tuning:

Tuning's impact on key performance metrics like Kappa, MCC, ROC area, and F-Measure was barely noticeable across training sets. This suggests limited benefit for achieving satisfactory performance, especially given the resource demands of tuning. Also, while accuracy gains on training sets were observed, their inconsistency and absence on some test sets raise concerns about overfitting and generalization to unseen data. Ultimately, the time and resources needed for tuning might not be justifiable if the baseline performance is already acceptable.

Finally, whether tuning is ultimately worth the effort depends on a delicate balance between your specific needs and available resources. In this case, even slight accuracy gains hold significant value and therefore, tuning can provide positive rewards in this case. However, if my baseline performance already would have met expectations, then skipping the complexities of tuning might be the wiser choice.

Final Evaluation of Final Test Set: The final evaluation of final test set of the predictive analytics project's in real estate, leveraging the "Housing & Ocean Proximity" dataset, showed that tuning the Sequential Minimal Optimization (SMO) classifier with a polynomial kernel resulted in slight but consistent improvements in accuracy across training sets. While the performance

enhancements were modest, they were significant enough to suggest that tuning is valuable when precision is crucial. However, the impact on key metrics like Kappa, MCC, and ROC area was minimal, indicating limited overall benefit for substantial performance improvement. The decision to tune depends on the specific requirements and resources available, suggesting that in cases where baseline performance is satisfactory, the complex and resource-intensive tuning process might not be necessary. Tuning can offer benefits when minor accuracy gains are critical, but its necessity is less clear when baseline performance meets expectations.

Also, to generate a Final result, train and test were combined into a single training set. The testing set was holdout, which had been other- wise unused. Using SMO with C = 0:15, 300 of 1000 instances were correctly classied, for a Kappa statistic of Using 1R, 300 instances were correctly clas-with accuracy **0.92**, for a Kappa statistic of **0.7731**.

My takeaway:

This work has offered a valuable learning experience for a student like me, interested in the intersection of predictive analytics, real estate, and machine learning. By diving into this analysis, I gained valuable insights into:

Unveiling Real Estate Trends with Machine Learning:

- Witnessed the power of machine learning to predict key indicators like housing prices and crime rates, empowering investors, urban planners, and policymakers.
- Understand the crucial role of integrating data on both house characteristics and neighborhood dynamics for accurate market forecasts, highlighting the importance of comprehensive data collection and analysis in real estate decision-making.

In term of mastering Machine Learning Techniques:

- Gained hands-on experience with Weka & LightSide and explored their suitability for specific classification and error analysis tasks within the real estate domain.

- Developed a deeper understanding of model evaluation metrics like Kappa statistic, confusion matrix, and feature weights, equipping you with the tools to assess and refine machine learning models effectively.

Limitations:

- Recognized the limitations of real-world models, including potential bias and data imbalances, fostering responsible data science practices and ethical considerations.

The model showed promising performance in predicting ocean proximity, but error analysis revealed areas for improvement, such as addressing class imbalance and enhancing decision boundaries between classes. By implementing the suggested solutions and incorporating additional features, the model's accuracy and generalizability could be further enhanced. The study also touched on tuning methodology in Weka, highlighting the importance of adjusting parameters for optimal performance. Several screenshots were used, illustrating the error analysis provided helping in building upon the analysis in detail.

My insights: I feel like,

-It would be interesting to explore the impact of specific features on the model's predictions.

-Comparing the performance of different machine learning algorithms could provide valuable insights.

-Investigating the reasons behind misclassifications could lead to further model refinement.

_

<u>References</u>

- [1] Deloitte Middle East. (n.d.). How AI can enhance urban planning, asset management and investments. Deloitte. Retrieved from Deloitte
- [2] Journal of Big Data. (n.d.). Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. SpringerOpen.
- [3] Freeman, L., Cassola, A. & Cai, T. (2016) Displacement and gentrification in England and Wales: A quasi-experimental approach. Urban Studies, 53(13), 2797–2814. Available from: <u>https://doi.org/10.1177/004209801559812</u> Ghaffari, L., Klein, J.-L. & Baudin, W.A. (2018)

- [4] Toward a socially acceptable gentrification: A review of strategies and practices against displacement. Geography Compass, 12(2), e12355. Available from: <u>https://doi.org/10.1111/gec3.12355</u>
- [5] Glass, R. (1964) London: Aspects of change. London: MacGibbon & Kee. Hamnett, C. (2003) Gentrification and the middle-class remaking of inner London, 1961–2001. Urban Studies, 40(12), 2401–2426. Available from: <u>https://doi.org/10.1080/0042098032000136138</u>
- [6] Case, Karl E. and Christopher J. Mayer, "Housing Price Dynamics within a Metropolitan Area," Regional Science and Urban Economics, 1996, 26 (3-4), 387–407.
- [7] Maryna Marynchenko, "Home Price Appreciation in Low-and Moderate-Income Markets," Low-income Homeownership: Examining the Unexamined Goal, 2002, pp. 239–256.