**TY9 Research Practices - CCE1 Literature Survey**

**Team code:** 1\_A  
**Team Member1:** Adarsh Kanaujiya/ 33 /9930410665  
**Team Member2:** Meet Katarmal / 34 / 9004054043   
**Team Member3:** Fatteh Ali / 26 / 70398 18429

**Tentative Title:**

Machine Learning Approaches for Real Estate Analytics: A Comparative Study

**Domain:**

Real Estate Analytics using AI/ML

**Sub Domain (for all 3):**

* Price Prediction (Regression/Ensemble Models)
* Real Estate Valuation via Image Analysis (CNNs, Computer Vision)
* Buyer Behavior & Demand Forecasting (NLP, Sentiment Analysis)

**Objective Description:**

To explore how Artificial Intelligence and Machine Learning techniques can transform real estate analytics. The study focuses on three major areas:

1. **Price Prediction** using regression, ensemble, and graph-based models to achieve accurate and dynamic property valuation.
2. **Image-based Valuation** of buildings and neighborhoods using CNNs and visual features to supplement pricing decisions.
3. **Buyer Behavior Analysis** through NLP and demand forecasting to capture market trends and preferences.

The aim is to compare traditional manual approaches with AI/ML-driven solutions to improve **accuracy, efficiency, and scalability** of real estate decision-making.

**PICO Section**

**Team Member1: Adarsh Kanaujiya**

**Subdomain: Price Prediction using AI/ML**

**PICO 1**

**PAPER TITLE:**

House Price Prediction using Machine Learning Models [Link of paper](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4108744)

**Authors of paper:**

Niharika Sharma, Abhinav Gupta, Shubham Soni (SSRN, 2022)

**Paper Description:**

* **Problem Statement:**  
  Accurately predicting house prices is a complex challenge due to the large number of influencing factors such as location, size, economic conditions, and neighborhood features. Traditional methods often fail to capture these nonlinear dependencies, leading to poor prediction accuracy. The paper addresses the problem of developing a reliable and scalable machine learning model to improve house price prediction.
* **Intervention:**  
  The authors applied **multiple machine learning algorithms** such as Linear Regression, Decision Trees, Random Forest, and Gradient Boosting to predict house prices. They used historical housing data containing various features (location, number of bedrooms, square footage, etc.) and trained models to establish relationships between features and price outcomes.
* **Comparison:**  
  The study compared the performance of different models, especially **Linear Regression vs. advanced ensemble methods** like Random Forest and Gradient Boosting. The goal was to determine which approach delivers the most accurate predictions with lower error rates.
* **Outcome:**  
  The results showed that **ensemble models (Random Forest & Gradient Boosting)** significantly outperformed Linear Regression in terms of prediction accuracy and robustness. These models captured complex, nonlinear relationships between housing features and price more effectively, proving to be more reliable tools for real estate price forecasting.

**PICO 2**

**PAPER TITLE:**  
*Using Ensemble Methods of Machine Learning to Predict Real Estate Prices* [*Link of paper*](https://arxiv.org/pdf/2504.04303)

**Authors of paper:**  
Oleh Pastukh and Viktor Khomyshyn, Ternopil Ivan Puluj National Technical University, Ukraine (2024, CEUR Workshop Proceedings)

**Paper Description:**

* **Problem Statement (P):**  
  Accurate forecasting of real estate prices is essential for buyers, sellers, and investors to reduce risks and make informed decisions. However, traditional appraisal and regression-based models often fail to capture the nonlinear and complex interactions among property features (location, amenities, size, etc.), leading to errors and inconsistencies.
* **Intervention (I):**  
  The authors applied **ensemble machine learning methods** such as Gradient Boosting Regressor, Extra Trees Regressor, HistGradient Boosting Regressor, and Random Forest Regressor to predict real estate values. The models were trained and evaluated using datasets of property listings.
* **Comparison (C):**  
  The study compared the performance of different ensemble models against one another, focusing on how well each handled nonlinearity and variance in real estate data. Unlike traditional regression methods, the ensemble approaches aim to reduce overfitting and improve generalization.
* **Outcome (O):**  
  The results showed that **Gradient Boosting Regressor achieved the highest accuracy**, while Extra Trees, HistGradient Boosting, and Random Forest also performed strongly. The models were evaluated using R², RMSE, and MAE, confirming that ensemble methods significantly outperform conventional regression-based approaches for real estate price prediction. The study suggests further improvements through preprocessing and anomaly detection.

**PICO 3**

**PAPER TITLE:**  
*An Optimal House Price Prediction Algorithm: XGBoost* [*Link of paper*](https://arxiv.org/pdf/2402.04082)

**Authors of paper:**  
Hemlata Sharma, Hitesh Harsora (Sheffield Hallam University, UK), and Bayode Ogunleye (University of Brighton, UK)

**Paper Description:**

* **Problem Statement (P):**  
  Accurate prediction of house prices is challenging because prices are influenced not only by property attributes (e.g., size, rooms) but also by neighborhood features and socioeconomic factors. Traditional regression-based models often fail to capture these nonlinear relationships, leading to poor predictive performance.
* **Intervention (I):**  
  The authors applied **XGBoost (Extreme Gradient Boosting)** along with other machine learning algorithms such as Support Vector Regressor, Random Forest Regressor, Multilayer Perceptron (MLP), and Multiple Linear Regression. They trained these models on the Ames City (Iowa, USA) housing dataset with feature engineering and hyperparameter tuning to optimize performance.
* **Comparison (C):**  
  The study compared **XGBoost** against SVR, Random Forest, MLP, and Linear Regression in terms of prediction accuracy and error reduction. Feature importance analysis was also performed to interpret model decisions.
* **Outcome (O):**  
  The results showed that **XGBoost outperformed all other models**, delivering the highest predictive accuracy and lowest error rates. Moreover, the model identified the most important factors influencing prices (location, property size, neighborhood amenities). This approach enables stakeholders—buyers, developers, and policymakers—to make **data-driven, reliable real estate decisions**.

**PICO 4**

**PAPER TITLE:**  
*A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models*

**Authors of paper:**  
Yahan Fu (Loudoun School for Advanced Studies, Ashburn, Virginia, USA) [Link of Paper](https://www.researchgate.net/publication/383208991_A_Comparative_Study_of_House_Price_Prediction_Using_Linear_Regression_and_Random_Forest_Models)

**Paper Description:**

* **Problem Statement (P):**  
  Accurately predicting house prices is vital for stakeholders in real estate, finance, and urban planning. Traditional linear regression models, while interpretable, struggle to capture the complexity and nonlinear patterns inherent in real estate data, resulting in suboptimal predictions. Therefore, there is a need to systematically evaluate whether more advanced machine learning methods such as Random Forest can offer significantly improved performance.
* **Intervention (I):**  
  The study applies two modeling approaches:
  1. **Linear Regression (LR)** — as a conventional baseline.
  2. **Random Forest (RF)** — an ensemble learning method that handles nonlinearity and feature interactions more effectively.

Both models are trained on a Kaggle housing dataset featuring predictors such as building class (“grade”), living area size (“sqft\_living”), construction year, and land area, among others .

* **Comparison (C):**  
  The performance of LR and RF models is compared using key evaluation metrics:
  1. **RMSE (Root Mean Squared Error)**
  2. **R² (Coefficient of Determination)**
* **Outcome (O):**
  1. **Linear Regression:** RMSE = 214,472.76; R² = 0.70
  2. **Random Forest:** RMSE = 148,428.13; R² = 0.85

The results show a substantial improvement with Random Forest, demonstrating its superior accuracy and ability to capture complex patterns in house price data. Additionally, feature importance analysis highlighted that “grade” and “sqft\_living” were the most influential predictors—providing interpretable insights for stakeholders

**PICO 5**

**PAPER TITLE:**

*Intelligent Feature Selection Ensemble Model for Price Prediction in Real Estate Markets* [*Link*](https://www.mdpi.com/2227-9709/12/2/52)

**Authors of paper:**

Daniel Cristóbal Andrade-Girón, William Joel Marin-Rodriguez, Marcelo Gumercindo Zuñiga-Rojas (Universidad Nacional José Faustino Sánchez Carrión, Peru)

**Paper Description:**

* **Problem Statement (P):**  
  Traditional models for real estate price prediction face challenges due to **high-dimensional datasets** and **redundant or irrelevant features**, which can lead to overfitting, poor model generalization, and computational inefficiency. There’s a need for models that maintain high accuracy while ensuring interpretability and speed.
* **Intervention (I):**  
  The authors propose an ensemble framework combining methods such as **Stacking**, Gradient Boosting, Random Forest, AdaBoost, Extra Trees, Bagging, and Voting. They incorporate **feature selection techniques** like Recursive Feature Elimination (RFE), Random Forest Feature Selection (RF-FS), and Boruta to reduce dimensionality effectively .
* **Comparison (C):**  
  They compare:
  + Full-feature ensemble models (using all ~227 variables).
  + Reduced-feature versions using RFE or Boruta (down to 15–16 variables).
  + Various ensemble algorithms with and without feature selection.
* **Outcome (O):**
  + **Stacking (full-feature):** R² = 0.924, RMSE ≈ 23,100, MAE ≈ 14,090, Concordance Correlation Coefficient (CCC) = 0.960, runtime ≈ 67.23 s.
  + **Gradient Boosting (full-feature):** R² = 0.920, MAE ≈ 14,540, CCC = 0.958.
  + With **Boruta** feature selection:
    - The Stacking model maintained strong performance (R² ≈ 0.908, MAE ≈ 15,472, CCC ≈ 0.951) but with **67% reduced runtime** (~22 s).
  + **RFE**, on the other hand, sometimes increased MAE and decreased R² modestly compared to full-feature models.

The results demonstrate that **ensemble models, especially Stacking**, offer **state-of-the-art predictive accuracy and robustness**, and when combined with Boruta feature selection, can achieve similar performance much more efficiently

**PICO Section**

**Team Member2:Meet Katarmal**

**Subdomain: Image-Based Valuation**

**PICO 1**

**Paper Title:** *Machine Learning Approach to Residential Valuation: A Convolutional Neural Network Model for Geographic Variation* [*Link*](https://link.springer.com/article/10.1007/s00168-023-01212-7)

* **Problem (P):**  
  Traditional real estate valuation heavily depends on hedonic pricing models, which rely on structured attributes such as square footage, number of rooms, and location coordinates. However, these models fail to capture fine-grained *geographic and neighborhood-level visual cues*, like quality of streets, greenery, or property façade. This creates valuation gaps between “similar” properties located in different environments.
* **Intervention (I):**  
  The study introduces a convolutional neural network (CNN) trained on property images along with spatial/geographic data. The CNN extracts features that represent subtle location-dependent characteristics—such as neighborhood density, architectural style, and building design—providing a richer feature set for price prediction.
* **Comparison (C):**  
  Results are compared against traditional regression-based models and hedonic pricing methods that only use tabular property data (area, location coordinates, number of rooms, etc.).
* **Outcome (O):**  
  CNN-based models significantly outperform baseline regression models, especially in regions with high geographic variation. They provide a more accurate and automated method of property valuation, reducing over-reliance on manual appraisers.

**PICO 2**

**Paper Title:** *Real Estate Valuation with Multi-Source Image Fusion and Enhanced Machine Learning Pipeline* [*Link*](https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0321951&utm_source)

* **Problem (P):**  
  Real estate valuation often uses a single type of data—either satellite imagery, property photographs, or textual descriptions. This single-source approach misses out on complementary signals, e.g., interior conditions, neighborhood environment, and curb appeal, all of which impact property prices.
* **Intervention (I):**  
  The authors propose a **multi-source image fusion framework** that integrates street view, satellite imagery, and interior photos into one deep learning pipeline. By combining different perspectives, the model learns comprehensive features of properties and their surroundings.
* **Comparison (C):**  
  Compared with models using only one type of image input (e.g., only satellite or only street view) and also against hedonic regression methods.
* **Outcome (O):**  
  Fusion-based deep models produce the best prediction results, capturing both environmental and interior quality. This reduces prediction error and makes valuation more robust across diverse property types.

**PICO 3**

**Paper Title:** *Using Images as Covariates: Measuring Curb Appeal with Deep Learning* [*Link*](https://arxiv.org/abs/2403.19915?utm_source)

* **Problem (P):**  
  “Curb appeal”—the attractiveness of a property’s exterior—is known to influence buyer decisions and market prices, but it is difficult to measure and has historically been ignored in automated models. Manual scoring of curb appeal is subjective and inconsistent.
* **Intervention (I):**  
  Deep learning CNNs are applied to property exterior photos to automatically quantify curb appeal. Extracted features serve as new covariates in price prediction models, enabling a numeric representation of aesthetic appeal.
* **Comparison (C):**  
  Compared with baseline models that exclude visual features and with manual assessments of curb appeal done by real estate agents.
* **Outcome (O):**  
  Inclusion of deep image features improves explanatory power of pricing models, providing a systematic way to capture “visual charm” and reducing subjective bias in valuations.

**PICO 4**

**Paper Title:** *Location, Location, Location: Satellite Image-Based Real-Estate Appraisal* [*Link*](https://arxiv.org/abs/2006.11406?utm_source)

* **Problem (P):**  
  Property values are heavily influenced by *neighborhood context*—such as green cover, road connectivity, nearby infrastructure, and land use patterns. These cannot be captured by property attributes alone, causing incomplete appraisals.
* **Intervention (I):**  
  The study leverages **satellite imagery** analyzed with deep learning to extract location-specific features such as urban density, vegetation, proximity to roads, and public amenities. These features are then fed into price estimation models.
* **Comparison (C):**  
  Benchmarked against models that use only property-specific features (area, rooms, year built) without any contextual neighborhood data.
* **Outcome (O):**  
  The satellite image–based models greatly enhance accuracy, showing that environmental and locational signals play a major role in valuation. The approach automates what human appraisers often intuitively consider.

**PICO 5**

**Paper Title:** *Image-Based Appraisal of Real Estate Properties* [*Link*](https://arxiv.org/pdf/1611.09180v1)

* **Problem (P):**  
  Manual property appraisal is time-intensive, requires expert appraisers, and often suffers from human subjectivity and inconsistency. Automated appraisal systems relying only on structured datasets fail to assess property conditions and aesthetics.
* **Intervention (I):**  
  A deep CNN is trained on property images to directly predict appraisal values. The system integrates learned visual features such as design quality, maintenance status, and appearance to estimate prices automatically.
* **Comparison (C):**  
  Compared against human appraisers and traditional hedonic models.
* **Outcome (O):**  
  Automated image-based appraisal is faster, more scalable, and produces consistent valuations. It reduces dependency on subjective human input and offers near real-time property assessment.

**PICO 6**

**Paper Title:** *Visual Estimation of Building Condition with Patch-Level ConvNets* [*Link*](https://arxiv.org/abs/1804.10113?utm_source)

* **Problem (P):**  
  Evaluating the physical condition of buildings is crucial for price appraisal, renovation planning, and insurance. Human inspectors provide subjective, error-prone evaluations, and large-scale surveys are costly.
* **Intervention (I):**  
  Patch-level convolutional neural networks analyze building photos by splitting them into smaller patches. The model detects cracks, wall discoloration, and visible structural issues, producing an overall condition score.
* **Comparison (C):**  
  Compared against human surveyor assessments and holistic image models (without patch-level detail).
* **Outcome (O):**  
  Patch-level CNNs provide more objective, fine-grained condition assessments, scaling to large datasets with reduced human effort. These condition scores improve overall price prediction accuracy.

**PICO 7**

**Paper Title:** *Vision-Based Real Estate Price Estimation* [*Link*](https://arxiv.org/abs/1707.05489?utm_source)

* **Problem (P):**  
  Property prices are influenced by external and interior visual cues (design, material quality, finishing, lighting, aesthetics), yet traditional models only rely on structured data like square footage and location. This leads to incomplete valuations.
* **Intervention (I):**  
  A deep learning vision-based framework processes property photographs to directly estimate price. Features are automatically extracted, eliminating the need for manual tagging of visual attributes.
* **Comparison (C):**  
  Benchmarked against hedonic regression models and ML models using only tabular data.
* **Outcome (O):**  
  Vision-based models improve predictive accuracy and help identify latent visual factors affecting price. This allows for end-to-end valuation using only images.

**PICO 8**

**Paper Title:** *A Multi-Modal Deep Learning Based Approach for House Price Prediction* [*Link*](https://arxiv.org/abs/2409.05335?utm_source)

* **Problem (P):**  
  Relying only on one data source—either structured property data, images, or text—limits model performance. For example, images show aesthetics, text provides seller descriptions, and tabular data provides core metrics. Ignoring one weakens prediction.
* **Intervention (I):**  
  The study proposes a **multi-modal deep learning framework** that combines property data (area, rooms, location), textual descriptions, and property images into a unified model. This integration helps capture complementary signals from each modality.
* **Comparison (C):**  
  Compared against unimodal models (only tabular or only vision) and simpler machine learning baselines.
* **Outcome (O):**  
  Multi-modal models outperform unimodal ones, achieving state-of-the-art prediction accuracy. They highlight the importance of combining structured, visual, and textual information for realistic valuation.

**PICO Section**

**Team Member3: Fatteali**

**Subdomain: Buyer Behavior Analysis using AI/ML**

**PICO 1:**

**PAPER TITLE**

Local Housing Market Sentiments and Returns: Evidence from China.

[**https://link.springer.com/article/10.1007/s11146-022-09933-w**](https://link.springer.com/article/10.1007/s11146-022-09933-w)

**Authors of paper:**

Shulin Shen, Yiyi Zhao, and Jindong Pang

**Paper Description:**

* **Problem Statement (P):**  
  Conventional housing return models largely ignore the influence of buyer and seller sentiment, which is an important behavioral factor in real estate markets. As a result, predictions based solely on fundamentals often fail to capture short-term fluctuations and long-term dynamics, reducing forecast accuracy.
* **Intervention (I):**  
  The study constructs monthly local housing sentiment indices for 18 major Chinese cities using second-hand housing transaction data between 2016 and 2020. These indices are incorporated into predictive models to better account for behavioral drivers of housing prices.
* **Comparison (C):**  
  The sentiment-based models are evaluated against traditional housing return models that rely only on fundamental factors and exclude sentiment indicators. This comparison highlights the added value of incorporating behavioral measures into real estate forecasting.
* **Outcome (O):**

The results show that housing sentiment strongly predicts future returns, with evidence of short-term underreaction and long-term overreaction patterns. Sentiment effects also vary across cities depending on supply inelasticity, and a feedback loop between sentiment and housing prices is observed, demonstrating improved forecasting accuracy and deeper market insights.

**PICO 2**

**PAPER TITLE:**

Gauging Airbnb Review Sentiments and Critical Key-Topics by Small Area Estimation

<https://link.springer.com/article/10.1007/s10260-024-00764-y>

**Authors of paper:**  
Luca Frigau, Giulia Contu, Marco Ortu, Andrea Carta, et al.

**Paper Description:**

* **Problem Statement (P):**  
  Airbnb reviews are often biased toward positive feedback, leaving negative experiences underreported and overlooked. This makes it difficult for hosts and the platform to identify critical guest concerns. Traditional aggregated analyses smooth out these nuances, limiting insights into location-specific or host-related issues.
* **Intervention (I):**  
  The study applies Natural Language Processing (sentiment analysis and topic modeling) to Airbnb reviews, focusing on extracting hidden negative sentiments. These results are then integrated with a small-area estimation framework using Mixed-Effect Random Forest (MERF) models. This approach allows for fine-grained analysis across neighborhoods and host types in Rome.
* **Comparison (C):**  
  Unlike conventional review methods that rely on aggregate-level averages, this framework accounts for geographic and host-specific differences. Traditional models often miss critical localized patterns and underrepresent negative feedback, whereas the proposed method emphasizes spatial and contextual detail.
* **Outcome (O):**

The findings reveal geographically nuanced sentiment patterns and highlight critical guest concerns such as cleanliness, communication, and value-for-money. This allows both hosts and the platform to take targeted actions to improve service quality. Compared to standard models, the approach captures more meaningful insights by focusing on local and host-level variations.

**PICO 3**

**PAPER TITLE:**

Can Social Media Information Amplify Short-term Housing Price Changes? An Investigation in China’s Major Cities

<https://link.springer.com/article/10.1007/s11518-024-5620-1>

**Authors of paper:**

Jieyun Wei, Xun Zhang, Zhiqiang Jiang, & Zhongjun Jin

**Paper Description:**

* **Problem Statement (P):**  
  Housing price fluctuations in China’s major cities are difficult to explain fully using only traditional economic fundamentals, such as supply-demand balance, interest rates, and income levels. There is growing uncertainty about whether public sentiment and expectations shared through social media platforms, like Sina Weibo, amplify short-term housing price dynamics.
* **Intervention (I):**  
  This study integrates a lexicon-based sentiment index derived from millions of Sina Weibo posts into panel econometric models of housing prices. By quantifying online optimism and pessimism, the intervention aims to capture demand-side psychological factors that conventional models often ignore.
* **Comparison (C):**  
  The performance of these enriched models is compared against baseline housing price models that rely solely on economic fundamentals, excluding social media sentiment
* **Outcome (O):**

Results show that social media sentiment significantly influences housing prices with a lagged effect. Specifically, a 0.1 increase in the sentiment index is associated with approximately a 0.2% rise in housing prices after two months. This finding confirms that optimistic expectations expressed online amplify short-term housing price increases, highlighting the predictive value of social media data in real estate analytics.

**PICO 4**

**PAPER TITLE:**

*Tracking Home-Owners’ Sentiments: Subjective Indices and Convergent Validity*

<https://link.springer.com/article/10.1007/s11146-023-09949-w>

**Authors of paper:**  
Sofie R. Waltl & Anthony Lepinteur

**Paper Description:**

* **Problem Statement (P):**  
  Conventional housing market models rely on transaction-based price indices, which may be unavailable or incomplete in some regions. Surveys that ask homeowners to estimate the current value of their property—referred to as Owner-Estimated Values (OEVs)—are underutilized due to skepticism about their reliability. The central challenge lies in determining whether these subjective estimates can validly capture real housing market dynamics.
* **Intervention (I):**  
  This study introduces rigorous methodologies to construct subjective house price indices using popular econometric methods (hedonic models, repeat-sales panels, ROA and hybrid indices). Leveraging survey microdata from sources like the American Housing Survey (AHS), Italian SHIW, and the European HFCS—covering multiple countries and decades—they derive OEV-based price indices and assess their fidelity.
* **Comparison (C):**  
  The newly constructed subjective price indices are benchmarked against traditional objective price indices derived from transaction data (O-RPPIs), which serve as the gold-standard—such as indices produced by BIS, central banks, and academic sources—over the same geographies and time frames.
* **Outcome (O):**

The findings demonstrate strong **convergent validity**: subjective indices constructed from OEVs closely track the dynamics of objective indices (correlations often 0.8–0.96 in the U.S. and Italy), indicating that homeowners collectively capture market trends accurately. However, subjective indices tend to systematically overestimate price *levels*, reflecting psychological biases like the endowment effect. Thus, OEV-based indices are reliable for tracking *price changes* but less reliable for measuring exact *price levels*

**PICO 5**

**PAPER TITLE:**

The Influence of Home Buyer Sentiment on Chinese Housing Prices—Based on Media Text Mining

<https://ccsenet.org/journal/index.php/ijef>

**Authors of paper:**  
Qing Gao & Tianxiao Zhao

**Paper Description:**

* **Problem Statement (P):**  
  In China’s major cities, housing prices have remained persistently high despite fluctuations in traditional macroeconomic indicators. It remains unclear to what extent **home buyer sentiment**—as shaped by media narratives—contributes to these sustained price levels beyond fundamentals like income growth, supply-demand fundamentals, or monetary policy.
* **Intervention (I):**  
  This study constructs a **Buyer Confidence Index** derived from media coverage. It analyzes 115,139 housing-related media articles using **text mining techniques** to extract sentiment and investor/buyer expectations. The derived index is then incorporated into **multiple linear regression models** to evaluate its explanatory power over housing prices (using Guangzhou as the case study).
* **Comparison (C):**  
  The models including the Buyer Confidence Index are compared against **baseline regression models** that use only traditional housing price determinants—without media-based sentiment variables.
* **Outcome (O):**

The inclusion of media-derived buyer sentiment significantly enhances explanatory power: the Buyer Confidence Index is **positively and significantly correlated** with housing prices, indicating the meaningful role of media-influenced buyer sentiment in shaping price dynamics