**TY9 Research Practices - CCE1 Literature Survey**

**Team code:** 1\_A  
**Team Member1:** Adarsh Kanaujiya/ 33 /9930410665  
**Team Member2:** Meet/[Roll No]/[Mobile]  
**Team Member3:** Fatteh/[Roll No]/[Mobile]

**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