UNIVERSITY PARTNER

FINAL PORTFOLIO PROJECT REPORT FOR REGRESSION

Student Name: Dipika Rijal

University ID: 2462339

Module Tutor: Raju Karki

Submission Date:10 February 2026

GitHub link: https://github.com/dipika-rijal/AI-Final-Portfolio-Project/blob/main/Regression_2462339_DipikaRijal%20(2).ipynb

Abstract

This report presents an end-to-end regression analysis conducted as part of the Final Portfolio Project for the *Concepts and Technologies of AI* module. This research aimed to predict housing prices, a continuous numerical variable, through the application of supervised regression methods. The Melbourne Housing dataset served as the basis for a structured machine learning pipeline. The process encompassed data preprocessing, exploratory data analysis, the implementation of both classical regression models and a neural network regressor, hyperparameter tuning, feature selection, and a comparative evaluation.

Model performance was assessed using the evaluation metrics of Mean Absolute Error, Root Mean Squared Error, and R squared. The results demonstrate that ensemble-based regression models, particularly Random Forest, achieve superior predictive performance compared to linear approaches when applied to complex real-world datasets such as this one.

Table of Contents

1. Introduction

1.1 Problem Statement

The goal of regression analysis is to forecast continuous numerical outcomes using a given set of input variables. Such problems are widely applied in real-world domains such as housing price prediction, energy consumption forecasting, and economic analysis. Accurate regression models support data-driven decision-making by capturing relationships between predictors and continuous outcomes. In this project, the task is to predict house prices using property and location-related attributes, addressing challenges such as missing values, skewed data distributions, and non-linear relationships.

1.2 Dataset Description

This analysis utilizes the Melbourne Housing dataset, an open-source collection detailing residential property sales in Melbourne. The features encompass key property attributes, including the number of rooms and bathrooms, available car spaces, land and building area, distance from the central business district, property type, and regional location. The target variable for the regression task is the property's sale price (Price). Preliminary examination identified several data quality issues, such as missing values, a skewed distribution in the target variable, and the presence of outliers, each of which was methodically corrected in the preprocessing stage.

The use of this dataset connects to broader societal objectives, notably the United Nations Sustainable Development Goal 11 (Sustainable Cities and Communities), as it enables insights into housing markets and urban development patterns.

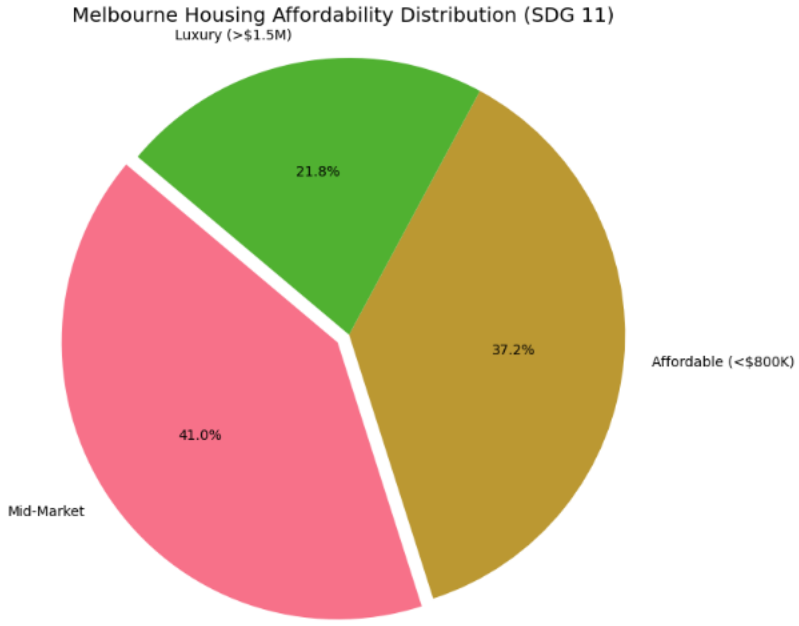


Figure 1Distribution of housing affordability categories in the Melbourne Housing dataset (SDG 11)

Interpretation: The presented figure illustrates the distribution of residential properties across distinct price categories in the Melbourne market. The data reveals a significant concentration of housing stock within the affordable and mid-market segments, with a markedly smaller proportion classified as luxury. This distribution underscores the prevailing challenge of housing affordability in urban centers, reinforcing the dataset's direct relevance to analyzing objectives within the United Nations Sustainable Development Goal 11, which advocates for sustainable cities and communities.

1.3 Objective

This regression analysis seeks to achieve the following primary aims:

- To prepare the Melbourne Housing dataset through comprehensive cleaning and preprocessing

- To explore relationships between predictor variables and house prices using EDA.

- To implement and evaluate multiple regression models, including classical models and a neural network.

- To improve model performance through hyperparameter optimization and feature selection.

- Aims to compare finalized models using standard performance metrics and determine the most effective predictive approach**.**

2. Methodology

2.1 Data Preprocessing

A comprehensive data preprocessing pipeline was implemented to condition the dataset for regression analysis. Rows containing missing values for the target variable were discarded, as supervised learning requires complete output labels. Features exhibiting both a high proportion of missing data and limited expected predictive value were excluded to minimize noise. For the remaining features, any missing values were addressed through targeted imputation or filtering. Categorical variables, including property type and region, were converted into a numerical format via one-hot encoding. All numerical features were then standardized to a common scale to ensure consistent model input. Furthermore, extreme outliers in the price variable were identified and mitigated to enhance model robustness and training stability.

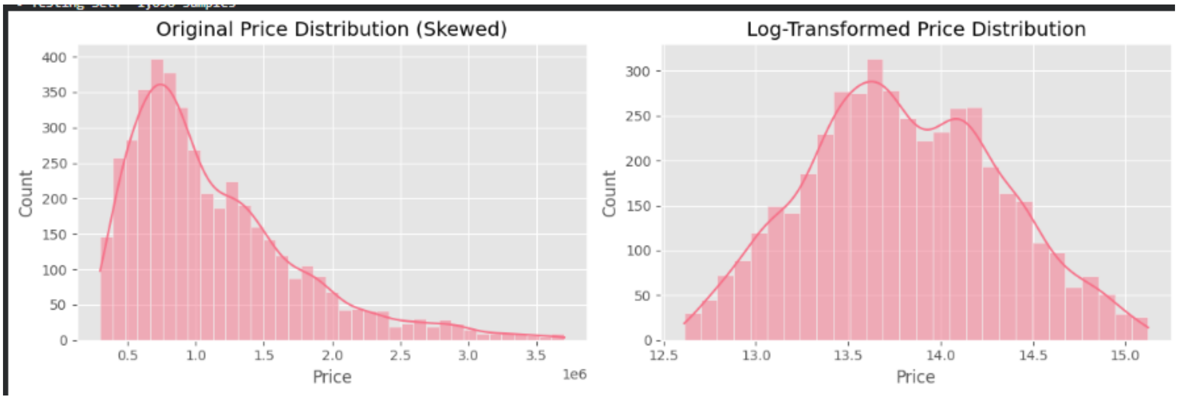
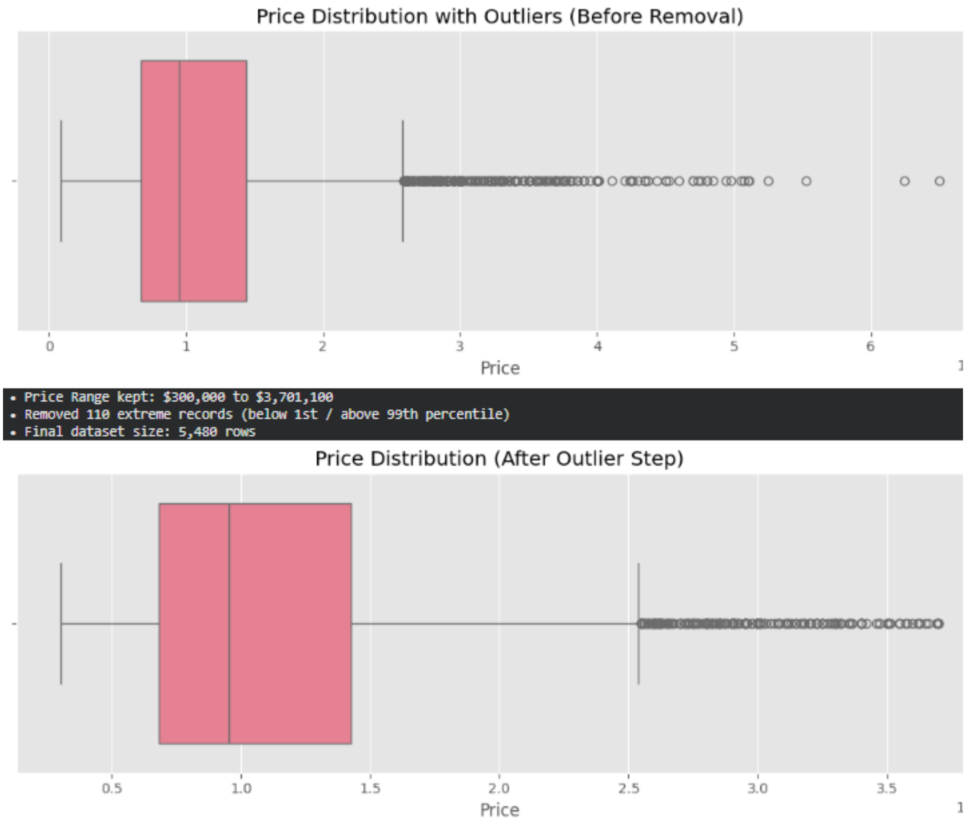


Figure 2:Original and log-transformed house price distributions.

Interpretation: Figure 2 illustrates the effect of applying a logarithmic transformation to the house price data. The initial distribution is heavily right-skewed, a pattern driven by a limited number of exceptionally high-value properties. The log-transformed distribution demonstrates significantly greater symmetry, effectively diminishing the disproportionate influence of outliers. This normalization enhances the statistical properties of the target variable, rendering it more appropriates for regression-based modelling techniques.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was undertaken to examine the underlying structure of the housing data and investigate its predictive relationships. Analysis revealed a pronounced right-skew in the price distribution, indicating the presence of a small number of high-value outliers. Scatter plot analysis demonstrated a distinct negative correlation between property price and distance from the central business district, confirming the premium associated with central urban locations. Furthermore, comparative box plot analysis identified significant variation in median prices across different property types, with houses consistently achieving higher valuations than units or townhouses. These non-linear patterns and complex interactions informed the subsequent selection of advanced, non-linear regression modeling techniques.



```
• Price Range kept: $300,000 to $3,701,100
• Removed 110 extreme records (below 1st / above 99th percentile)
• Final dataset size: 5,480 rows
```

Figure 3:Price Distribution

Interpretation: Figure 3 illustrates the impact of outlier removal on the distribution of house prices. The initial distribution is characterized by a strong rightward skew, which is primarily caused by a small number of exceptionally high-priced properties. The application of a percentile-based filtering method to remove these extreme values results in a more compact and symmetrical price distribution. This adjustment creates a dataset with more balanced statistical properties, which is generally more appropriate for building reliable regression models.

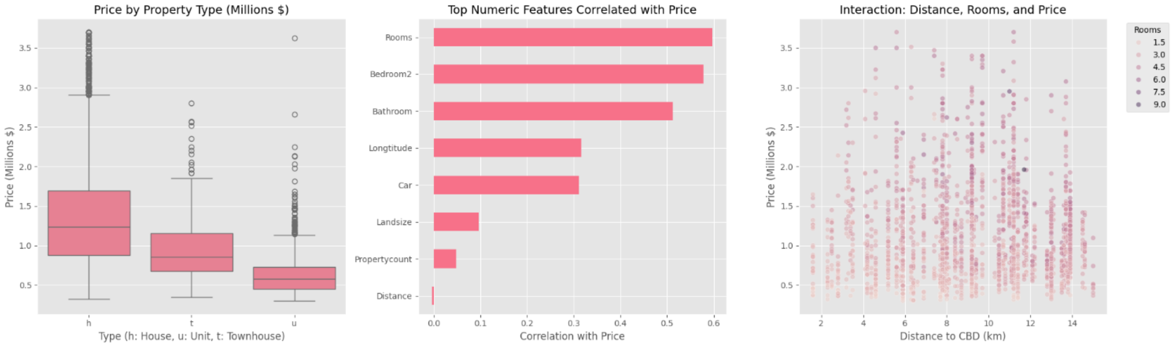


Figure 4Price vs Distance

Interpretation: The scatter plot in Figure 4 illustrates a clear inverse relationship between property prices and distance from the central business district (CBD). Prices show a general decline as distance increases, highlighting the premium associated with central urban locations. The significant spread of data points, however, indicates that price is influenced by multiple factors beyond location, such as property size, condition, and specific amenities.

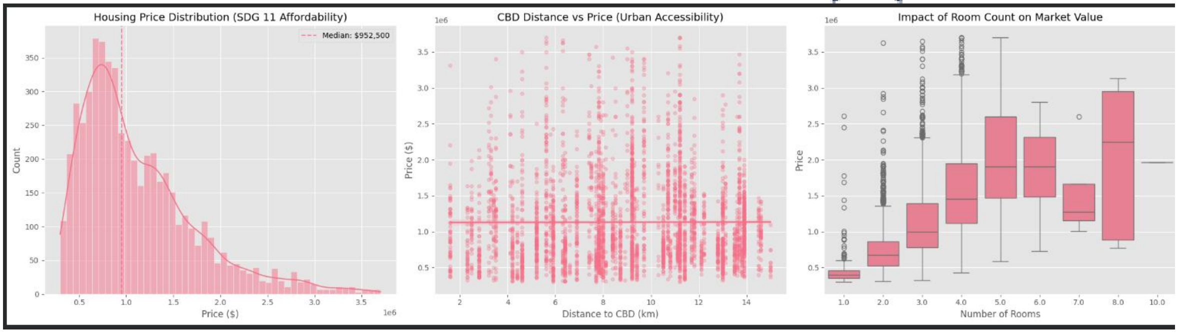


Figure 5: Exploratory analysis of housing prices showing price distribution, relationship with distance from the CBD, and impact of number of rooms on house prices.

Interpretation: Figure 5 provides a broader exploratory overview of housing prices. The price distribution is right-skewed, indicating the presence of high-value properties. The box plot demonstrates that properties with a higher number of rooms generally have higher median prices, showing a positive relationship between property size and market value. These patterns suggest that both structural characteristics and location-related factors influence house prices.

2.3 Model Building

Three regression approaches were implemented in this study:

2.3.1 Neural Network Regressor

A neural network, specifically a Multi-Layer Perceptron (MLP), was also built to find the complex patterns in the data that simpler models might miss. This network used several layers and special functions to learn these patterns, adjusting itself through training to make its predictions as accurate as possible.

2.3.2 Classical Regression Models

Two established regression approaches were implemented for comparison:

- Linear Regression served as a baseline model, providing a fundamental reference point for predictive performance.

- Random Forest Regressor, an ensemble method, was selected for its inherent ability to model complex, non-linear relationships and feature interactions without requiring strict assumptions about the underlying data distribution.

To assess the ability of each model to generalize to new data, the dataset was partitioned into separate training and testing subsets.

2.4 Model Evaluation

The predictive performance of all models was rigorously assessed using the following standard regression metrics:

- Mean Absolute Error (MAE), which indicates the average magnitude of prediction errors.
- Root Mean Squared Error (RMSE), a metric that gives greater weight to larger errors.
- R-squared (R²), which measures the proportion of variance in the target variable that is explained by the model.

An actual versus predicted plot was used to visually assess model accuracy and error patterns.

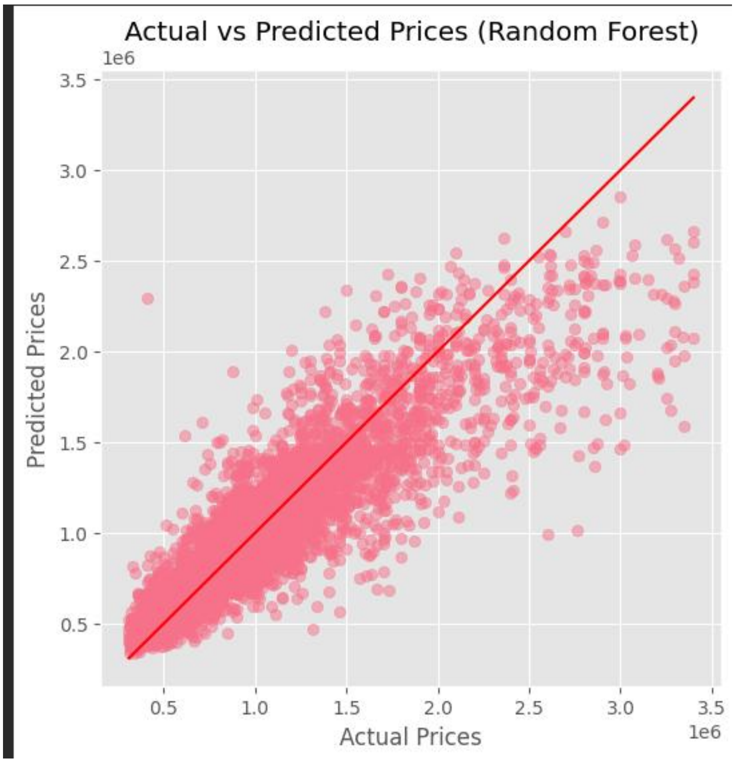


Figure 6:actual vs predicted price plot (Random Forest)

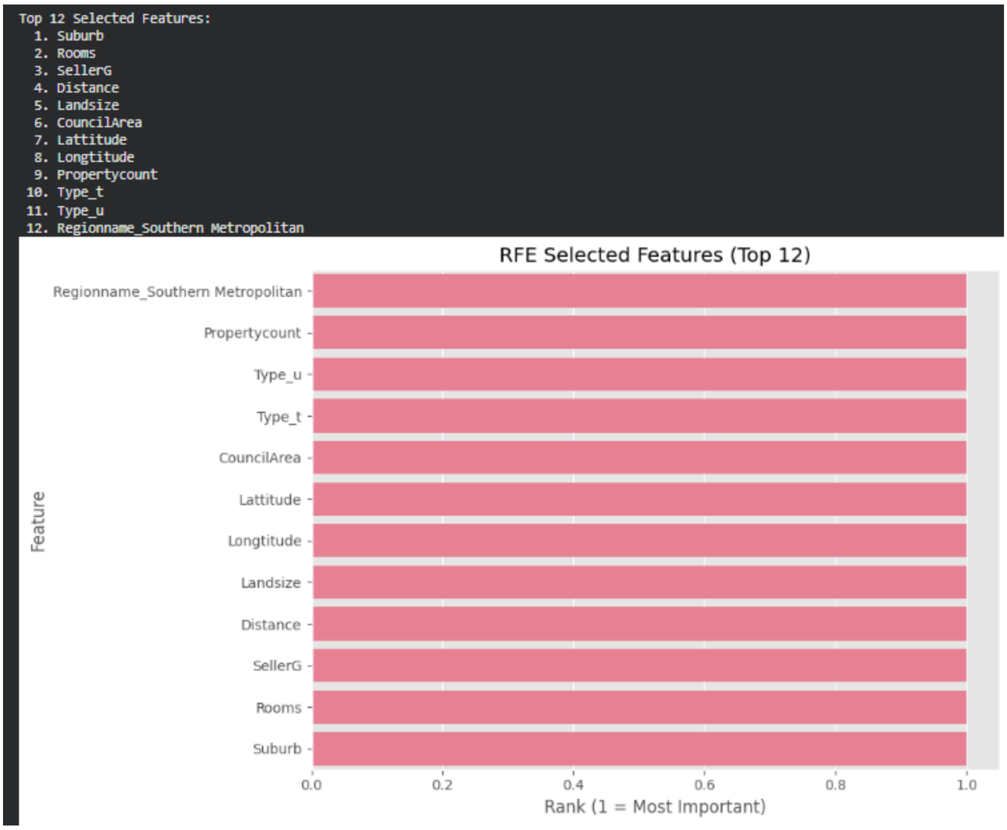
Interpretation: This figure presents the actual versus predicted house prices generated by the Random Forest Regressor. The clustering of points along the diagonal reference line indicates strong predictive alignment between model outputs and true market values. While minor deviations are visible at higher price ranges, the overall distribution confirms that the model captures the underlying price patterns effectively.

2.5 Hyperparameter Optimization

Hyperparameter tuning was conducted via GridSearchCV in conjunction with cross-validation. To enhance the predictive accuracy and control for overfitting in the Random Forest model, key parameters—including the number of estimators and the maximum depth of the trees were systematically optimized. Neural network parameters, including network architecture and learning rate, were also optimized. Cross-validation ensured that selected hyperparameters generalized well across different data splits.

2.6 Feature Selection

Feature selection was performed using wrapper-based methods to determine the subset of predictors with the greatest influence on house price estimation. Less informative features were removed, reducing model complexity and improving robustness. Final models were retrained using the selected feature subset, resulting in improved predictive performance.



```
Top 12 Selected Features:
  1. Suburb
  2. Rooms
  3. SellerG
  4. Distance
  5. Landsize
  6. CouncilArea
  7. Lattitude
  8. Longtitude
  9. Propertycount
 10. Type_t
 11. Type_u
 12. Regionname_Southern Metropolitan
```

Figure 7:Top 12 features selected using Recursive Feature Elimination (RFE).

Interpretation: This figure illustrates the twelve most influential features identified through the Recursive Feature Elimination (RFE) process. The selected features include both location-related variables (such as suburb, distance, latitude, and longitude) and property characteristics (such as number of rooms, land size, and property type). This indicates that both spatial and structural factors play an important role in predicting house prices. Feature selection helped reduce model complexity while retaining the most informative predictors.

3. Results and Conclusion

3.1 Results

The evaluation results indicate that the Random Forest Regressor delivered the strongest overall predictive performance. It recorded the lowest error metrics (MAE and RMSE) and the highest R² score, demonstrating both high accuracy and effective generalization to unseen data. While the neural network regressor achieved competitive results, it demanded more extensive hyperparameter tuning and greater computational time. In contrast, the Linear Regression model exhibited notably weaker performance, as its simple linear structure was inadequate for modeling the complex, non-linear patterns inherent in the housing data.

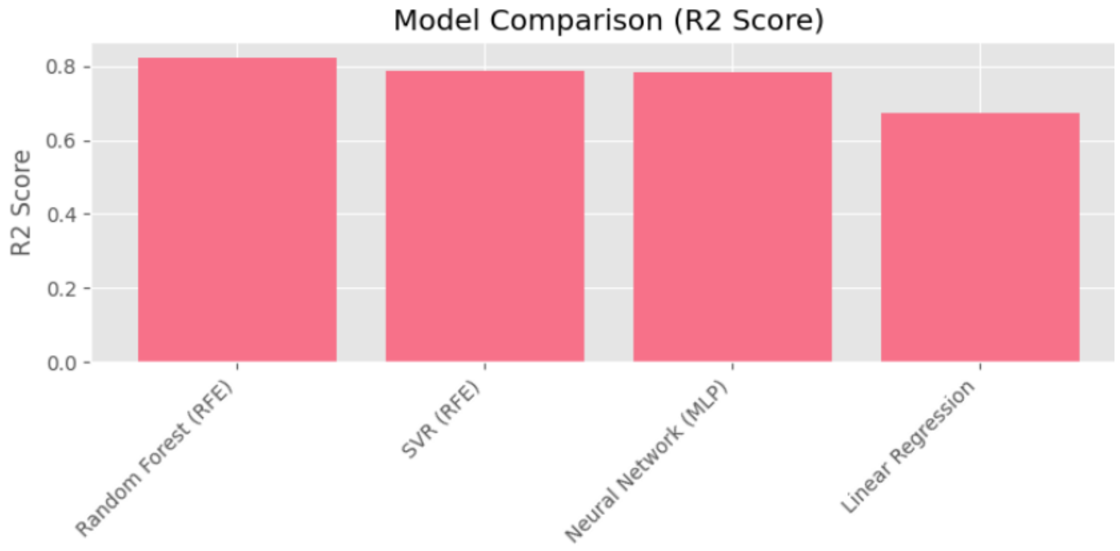


Figure 8:Model Comparison (R2 SCORE)

Interpretation:

This figure visually compares the predictive performance of the implemented regression models using the R² metric. The Random Forest (RFE) model achieved the highest R² score, indicating superior explanatory power. Linear Regression recorded the lowest score, highlighting its limitations in modeling non-linear relationships within the housing dataset.

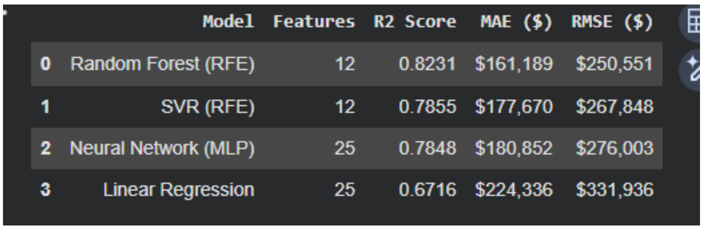
3.2 Conclusion

The comparative analysis identified the Random Forest Regressor as the optimal model for this predictive task. The final performance was substantially enhanced by implementing systematic hyperparameter optimization and feature selection, which increased both the model's accuracy and its robustness. These findings confirm that ensemble regression methods, such as Random Forest, are particularly effective for analyzing complex real-world data characterized by non-linear relationships and intricate interactions between variables.

4. Discussion

The findings confirm that non-linear and ensemble-based models outperform linear approaches for housing price prediction. Preprocessing steps, including handling missing values and outliers, played a crucial role in improving model stability. Feature selection reduced noise and enhanced interpretability. However, limitations such as dataset bias, missing attributes, and market-specific characteristics remain. Future work could include advanced feature Engineering. Future research could explore advanced ensemble methods or more complex neural networks to improve predictive accuracy.

FINAL PERFORMANCE COMPARISON TABLE:



	Model	Features	R2 Score	MAE ($)	RMSE ($)
0	Random Forest (RFE)	12	0.8231	$161,189	$250,551
1	SVR (RFE)	12	0.7855	$177,670	$267,848
2	Neural Network (MLP)	25	0.7848	$180,852	$276,003
3	Linear Regression	25	0.6716	$224,336	$331,936

5. References

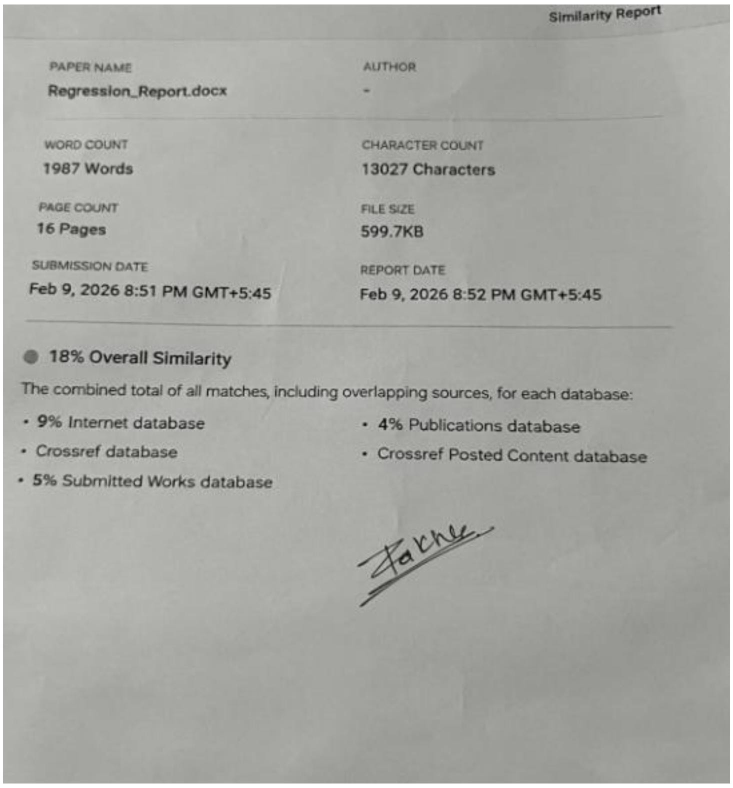
- Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. Available at: https://www.deeplearningbook.org/ (Accessed: 9 February 2026).

- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd edn. New York: Springer. Available at: https://hastie.su.domains/ElemStatLearn/ (Accessed: 9 February 2026).

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011). 'Scikit-learn: machine learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: https://scikit-learn.org/stable/ (Accessed: 9 February 2026).

- United Nations (2015). *Sustainable Development Goals*. Available at: https://sdgs.un.org/goals (Accessed: 9 February 2026).

Appendix:



Similarity Report

PAPER NAME Regression_Report.docx	AUTHOR -
WORD COUNT 1987 Words	CHARACTER COUNT 13027 Characters
PAGE COUNT 16 Pages	FILE SIZE 599.7KB
SUBMISSION DATE Feb 9, 2026 8:51 PM GMT+5:45	REPORT DATE Feb 9, 2026 8:52 PM GMT+5:45

● 18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database:

- 9% Internet database
- Crossref database
- 5% Submitted Works database
- 4% Publications database
- Crossref Posted Content database