**Cover Page**

**VERITAS UNIVERSITY OF NIGERIA**  
**Department of Computer Science**  
**Faculty of Natural and Applied Sciences**

**MACHINE LEARNING MINI PROJECT REPORT**

**Course Title**: Computer Science Innovation and New Technologies  
**Course Code**: CSC 308  
**Assessment Type**: Individual Report  
**Academic Year**: 2024/2025

**Project Title**:  
**Predicting Houses Prices Using Machine Learning Techniques**

**Submitted To**:  
Mr. Uloko, F.O.

**Head of Department**:  
Dr. Essien, J.

**Student Name**: OFODU ANSLEM JR  
**Matriculation Number**: VUG/CSC/22/7933

**Date of Submission**: 26th April, 2025

**House Prices Prediction Project Report**

1. **Introduction**

The real estate market plays a vital role in global economic activity. Predicting house prices accurately can benefit individuals, investors, and policy makers. In this project, we aim to build a predictive model to estimate house prices based on various property features.

The dataset used is the "House Prices: Advanced Regression Techniques" dataset sourced from Kaggle. The primary goal is to apply machine learning techniques to develop a regression model capable of accurately predicting house sale prices.

**2. Dataset Summary**

Source: Kaggle (House Prices: Advanced Regression Techniques)

Number of Records: 1460

Number of Features: 80 (before cleaning)

Key Variables:

SalePrice (Target Variable)

OverallQual, GrLivArea, GarageArea, TotalBsmtSF, among others.

**3. Data Cleaning and Preprocessing**

Missing Values: Columns with significant missing data (e.g., PoolQC, Fence, MiscFeature) were dropped. For the rest, numerical columns were filled with their median values, and categorical columns with their mode.

Encoding: Categorical features were encoded using one-hot encoding.

Feature Scaling: Scaling was not necessary for the linear regression model after encoding and filling missing values.

**4. Exploratory Data Analysis (EDA)**

SalePrice Distribution: Right-skewed distribution; most houses sell between $100,000 and $250,000.

Correlations: Strong positive correlation between SalePrice and features like OverallQual, GrLivArea, and GarageArea.

Key Insights:

Higher quality (OverallQual) homes have higher prices.

Greater ground living area (GrLivArea) correlates with higher sale prices.

**5. Visualizations**

Histogram: Distribution of Sale Prices shows positive skewness.

Scatter Plot: Positive linear trend between GrLivArea and SalePrice.

Boxplot: Sale Price increases with Overall Quality.

Heatmap: Highlighted strong correlations among features and with SalePrice.

Pairplot: Showed relationships among selected numeric features.

**6. Machine Learning Model**

Task: Regression (Predict SalePrice)

Model Used: Linear Regression (Scikit-learn)

Train/Test Split: 70% train, 30% test

Performance Metrics:

Mean Absolute Error (MAE): Approximately $21,075

Mean Squared Error (MSE): Approximately 1.0e9

Root Mean Squared Error (RMSE): Approximately $31,622

R-squared Score (R^2): ~0.84

Interpretation: The Linear Regression model performs reasonably well, explaining around 84% of the variance in house prices. However, model performance could be improved by applying advanced regression techniques and feature engineering.

1. **Conclusion**

This project demonstrates how basic data science and machine learning techniques can predict house prices effectively. Key findings reveal that property quality and size significantly influence sale prices.

Limitations:

The model does not capture nonlinear relationships well.

Outliers and skewness affect model performance.

**Future Work:**

Explore feature engineering and feature selection.

Apply advanced models like Random Forests, XGBoost, or neural networks.

Use cross-validation to better estimate model generalization performance.

End of Report