Indonesian Journal of Innovation Studies

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457



ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

Table Of Contents

Journal Cover 1
Author[s] Statement 3
Editorial Team 4
Article information 5
Check this article update (crossmark) 5
Check this article impact 5
Cite this article 5
Title page 6
Article Title 6
Author information 6
Abstract 6
Article content 8

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Copyright © Author(s). This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

EDITORIAL TEAM

Editor in Chief

Dr. Hindarto, Universitas Muhammadiyah Sidoarjo, Indonesia

Managing Editor

Mochammad Tanzil Multazam, Universitas Muhammadiyah Sidoarjo, Indonesia

Editors

Fika Megawati, Universitas Muhammadiyah Sidoarjo, Indonesia

Mahardika Darmawan Kusuma Wardana, Universitas Muhammadiyah Sidoarjo, Indonesia

Wiwit Wahyu Wijayanti, Universitas Muhammadiyah Sidoarjo, Indonesia

Farkhod Abdurakhmonov, Silk Road International Tourism University, Uzbekistan

Bobur Sobirov, Samarkand Institute of Economics and Service, Uzbekistan

Evi Rinata, Universitas Muhammadiyah Sidoarjo, Indonesia

M Faisal Amir, Universitas Muhammadiyah Sidoarjo, Indonesia

Dr. Hana Catur Wahyuni, Universitas Muhammadiyah Sidoarjo, Indonesia

Complete list of editorial team (link)

Complete list of indexing services for this journal (link)

How to submit to this journal (link)

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

Article information

Check this article update (crossmark)



Check this article impact (*)















Save this article to Mendeley



(*) Time for indexing process is various, depends on indexing database platform

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

Innovation In Smartphone Price Prediction Using Cross Validated Regression Models

Adiyanto Oktavianus Buaton, adiyantooktavianus7@gmail.com (*)
Computer Science, Faculty Of Engineering And Computer Science, Universitas Potensi Utama, Indonesia

Budi Triandi, buditriandi@gmail.com
Computer Science, Faculty Of Engineering And Computer Science, Universitas Potensi Utama, Indonesia

Hartono Hartono, hartono@staff.uma.ac.id
Computer Science, Faculty Of Engineering And Computer Science, Universitas Potensi Utama, Indonesia

(*) Corresponding author

Abstract

General Background The rapidly growing smartphone market necessitates data-driven pricing models to help consumers and manufacturers estimate valuations based on complex hardware specifications. **Specific Background** While machine learning algorithms like Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN) are widely used for product valuation, the unique heterogeneity and rapid cycles of the mobile market demand domain-specific evaluations. **Knowledge Gap** Existing empirical studies present mixed performance results and frequently neglect the generalization stability of these models under cross-validation scenarios. **Aims** This study evaluates linear SVM and K-NN models for predicting smartphone prices using technical attributes, explicitly analyzing the trade-off between predictive accuracy and model robustness. **Results** Utilizing a preprocessed dataset of 836 smartphone records, the K-NN model achieved lower test set errors with a Mean Absolute Error of 12,809.63 and an R^2 of 0.461, whereas the linear SVM recorded an R^2 of -0.013. However, 10-fold cross-validation exposed severe overfitting in the K-NN model with an R^2 of -0.659, while the SVM showed slightly better stability with an R^2 of 0.098. **Novelty** This research distinctly highlights the vulnerability of instance-based learning to local data density and skewed distributions, contrasting it with the constrained expressiveness of margin-based linear models in price prediction. **Implications** Future development of pricing recommendation systems must prioritize non-linear kernel exploration, robust hyperparameter optimization, and extensive cross-validation over simple error minimization.

Highlights

- K-Nearest Neighbors regression minimizes test set errors but demonstrates severe overfitting during cross-validation.
- Linear Support Vector Machine provides marginally higher generalization stability despite weaker single-split predictive accuracy.
- Right-skewed technical specification data necessitates robust validation schemes and non-linear kernel optimization.

Keywords

Smartphone Price Prediction; K-Nearest Neighbors; Support Vector Machine; Regression Analysis; Cross-Validation

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

Published date: 2026-11-04

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

INTRODUCTION

The rapid growth of the smartphone market has intensified competition among manufacturers to offer devices with increasingly sophisticated technical specifications at competitive prices. For consumers, assessing whether a smartphone's price is commensurate with its technical features remains challenging due to the complex and often non-linear relationships between hardware specifications and market valuation. As a result, data-driven price prediction models have gained attention as decision-support tools for consumers and manufacturers in estimating fair and competitive pricing [1].

Previous studies have applied various machine learning techniques for product price prediction, including Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN). SVM has been widely adopted due to its theoretical robustness in handling high-dimensional feature spaces and its applicability to regression tasks, while K-NN offers an intuitive instance-based approach that estimates prices based on similarity to previously observed products [2][3]. Although these methods have been explored in domains such as property and vehicle price prediction [4][5], their findings cannot be directly generalized to smartphone pricing, as the smartphone market is characterized by rapid product cycles, strong brand effects, and highly heterogeneous feature combinations that evolve over short time spans. This domain-specific heterogeneity motivates a focused investigation on smartphone price prediction rather than relying on conclusions drawn from other product categories.

Several comparative studies have reported mixed results regarding the relative performance of SVM and K-NN in regression tasks. While some works suggest that SVM can achieve superior accuracy under certain data distributions [6], others indicate that K-NN may perform better in capturing local non-linear patterns [7]. However, prior studies often differ in dataset characteristics, feature representations, and evaluation protocols, making their conclusions difficult to generalize. Moreover, limited attention has been paid to the stability and generalization of these models under cross-validation in the context of smartphone price prediction. This inconsistency in empirical findings, combined with the lack of domain-specific evaluation for smartphones, constitutes a research gap that this study aims to address [8].

In addition, existing studies frequently focus on achieving low prediction error on a single train–test split without sufficiently discussing model robustness across different data partitions. Given the relatively small size and potentially noisy nature of publicly available smartphone pricing datasets, model performance may be sensitive to sampling variation [9]. Furthermore, smartphone prices are influenced by market heterogeneity, including brand positioning, regional pricing strategies, and currency scales, which may affect the learning behavior of regression models. Normalization of price scales and feature distributions is therefore a critical preprocessing step that warrants explicit consideration in empirical evaluations [10].

This study focuses on comparing Linear Support Vector Machine (SVM) and K-Nearest Neighbors Regression (K-NN) as two representative yet methodologically contrasting approaches: a margin-based parametric model versus an instance-based non-parametric model. The selection of these methods is motivated by their widespread use, interpretability, and contrasting learning biases, which allow for a focused analysis of accuracy–stability trade-offs under limited data conditions [11]. While more complex ensemble methods such as Random Forest or Gradient Boosting have demonstrated strong performance in price prediction tasks, this study deliberately restricts the scope to SVM and K-NN to provide a controlled baseline comparison and to analyze how simpler models behave under cross-validation when applied to heterogeneous smartphone feature data [12].

Using a dataset of smartphone technical specifications, this research evaluates model performance through 10-fold cross-validation and standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score [13]. By emphasizing both predictive accuracy and generalization stability, this study seeks to provide empirical insights into the practical suitability of SVM and K-NN for smartphone price prediction. The findings are expected to inform the design of data-driven pricing recommendation systems and to offer guidance on model selection under data and resource constraints [14].

METHOD

This study adopts a quantitative experimental design to evaluate the performance of two regression-based machine learning models, namely Support Vector Machine (SVM) and K-Nearest Neighbors Regression (K-NN), for predicting smartphone prices based on technical specifications. The research workflow consists of data collection, data preprocessing, model configuration and hyperparameter tuning, model training, performance evaluation, and result interpretation [15].

A. Data Collection

The dataset used in this study is obtained from Kaggle, specifically the Mobile Price Prediction Dataset published by JsonAli2003 (https://www.kaggle.com/datasets/jsonali2003/mobile-price-prediction-dataset/data). The dataset contains 836 smartphone records with the following attributes: Brand, Model, Color, RAM, Storage, Primary Camera, Selfie Camera, Battery capacity, Processor, Screen Size, and Price as the target variable.

B. Data Preprocessing

Several preprocessing steps are applied to ensure data quality and compatibility with machine learning algorithms :

1.Duplicate entries are removed to avoid bias in model learning.

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

2. Missing values, if present, are handled using appropriate imputation strategies.

3. Nominal categorical variables, including Brand, Model, Color, and Processor, are transformed using One-Hot Encoding to prevent the introduction of artificial ordinal relationships, which is particularly important for distance-based models such as K-NN.

4. Numerical features and the target variable are separated, with Price defined as the regression target.

5. Feature scaling is performed using StandardScaler to normalize numerical features to zero mean and unit variance. Standardization is selected due to the sensitivity of SVM to feature scales and to ensure fair distance computation in K-NN. While alternative scaling methods such as Min–Max normalization may be suitable for skewed distributions, StandardScaler is adopted to maintain comparability across features and models.

6. Exploratory feature relevance analysis is conducted using correlation analysis and mutual information scores to provide preliminary insights into the relative contribution of each technical attribute to price variation [16].

C. Splitting Training and Testing Data

To ensure a consistent and unbiased evaluation framework, this study employs a two-level evaluation strategy. The dataset is first partitioned into training (80%) and testing (20%) subsets using a fixed random seed (random_state = 42) to enable reproducibility. Model selection and hyperparameter tuning are conducted exclusively on the training set using 10-fold cross-validation. The final model configurations obtained from cross-validation are then evaluated on the held-out testing set to assess generalization performance. This protocol prevents information leakage and ensures consistency between validation and final testing results.

D. Machine Learning Model Development

Two machine learning models are implemented in this study:

1. Support Vector Machine (SVM)

SVR is implemented with multiple kernel functions to capture both linear and non-linear relationships between smartphone features and prices. The kernel options include linear, radial basis function (RBF), and polynomial kernels. Hyperparameters such as the regularization parameter (C), kernel-specific parameters (e.g., gamma for RBF), and epsilon in the ε-insensitive loss function are optimized using grid search with cross-validation. This configuration aims to mitigate underfitting and improve the model's ability to capture complex pricing patterns.

2.K-Nearest Neighbors Regression (K-NN)

K-NN regression is configured with varying numbers of neighbors ($k \in \{3, 5, 7, 9, 11\}$) and distance metrics. A grid search procedure is employed to identify the optimal value of k that minimizes validation error. This sensitivity analysis addresses the arbitrariness of selecting a fixed k value and provides a more robust configuration for price prediction. The computational complexity of K-NN, which increases with dataset size due to its instance-based nature, is also considered in evaluating the model's practicality for real-world deployment [17].

E. Model Evaluation

Model performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score. Cross-validation results on the training set are used for model selection and hyperparameter tuning, while the final performance is reported on the independent testing set to assess real-world generalization. This separation ensures that evaluation metrics reflect unbiased predictive performance rather than optimistic estimates derived from resubstitution or single-split validation.

F. Prediction Result

The optimized SVM and K-NN models are applied to the testing dataset to generate price predictions. Predicted prices are compared with actual prices to assess model accuracy and error patterns. The results are summarized in tabular form, including actual prices, predicted prices from each model, and corresponding smartphone brands for interpretability. Additionally, error distributions are analyzed to identify systematic biases in prediction, such as underestimation or overestimation in specific price ranges [18].

RESULTS AND DISCUSSION

A. Data Collection

The "Mobile Price Prediction Dataset" used in this study consists of 836 records with various features representing smartphone technical specifications, including Brand, Model, 4G, 5G, Battery Capacity, Internal Memory, RAM, Primary Camera, Secondary Camera, Processor Speed, Screen Size, Weight, and Price as the target variable.

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

After data cleaning, which includes checking for duplicates and handling missing values, the dataset is transformed into a numerical representation and standardized to ensure comparable feature scales across models. The dataset is then divided into 80% training data (669 records) and 20% testing data (167 records) to evaluate the performance of Support Vector Machine (SVM) and K-Nearest Neighbors Regression (K-NN) in predicting smartphone prices.

Exploratory analysis is conducted using data visualization to examine feature distributions and potential relationships with price. The price distribution exhibits a right-skewed pattern, indicating dominance of low-priced smartphones and relatively sparse high-priced samples. This imbalance is important because it may bias regression models toward the majority price range [19]. Figure 1 illustrates the overall structure of the dataset.

Dataset Asli:

	Unnamed: 0	Brand ma	Ratings	RAM	ROM	Mobile_Size	Primary_Cam	Selfi_Cam	Battery_Power	Price
0	0	LG V30+ (Black, 128)	4.3	4.0	128.0	6.00	48	13.0	4000	24999
1	1	I Kall K11	3.4	6.0	64.0	4.50	48	12.0	4000	15999
2	2	Nokia 105 ss	4.3	4.0	4.0	4.50	64	16.0	4000	15000
3	3	Samsung Galaxy A50 (White, 64)	4.4	6.0	64.0	6.40	48	15.0	3600	18999
4	4	POCO F1 (Steel Blue, 128)	4.5	6.0	128.0	6.18	35	15.0	3800	18999
...	...	...	...	...	...	...	...	...	...	...
831	831	Karbonn K24 Plus Pro	3.8	6.0	32.0	4.54	48	12.0	2000	1299
832	832	InFocus POWER 2	4.1	8.0	64.0	4.54	64	NaN	2500	1399
833	833	Alcatel 5V (Spectrum Blue, 32)	4.4	3.0	32.0	6.26	48	1.0	3800	9799
834	834	JIVI JV 12M	3.7	10.0	32.0	4.50	64	NaN	3500	799
835	835	Blaoar BS Grip	3.5	6.0	32.0	4.50	64	15.0	1050	700

Figure 1. *Mobile Price Prediction Dataset*

B.Data Preprocessing Result

The preprocessing stage aims to ensure data quality and compatibility with machine learning models. The steps and outcomes are as follows:

1. Removal of Duplicates and Missing Data

a. The dataset contains 836 records after cleaning.

b. No duplicate records are detected.

c. Missing values, if present, are handled to avoid bias in model training.

2.Conversion of Categorical Data to Numeric

a. The “Brand”, “Model”, and “Processor” attributes are categorical.

b. These attributes are converted into numerical representations to enable model training.

c. It is noted that the encoding scheme may introduce implicit assumptions about category relationships; therefore, its potential impact on distance-based learning (K-NN) is acknowledged as a methodological limitation.

3. Feature and Target Separation

a. The target variable is defined as “Price”.

b. The feature set includes technical specifications such as RAM, Internal Memory, Camera resolution, Battery capacity, Processor speed, Screen size, network support, and weight.

4. Splitting Data into Training and Testing Sets

a. The dataset is split into 80% training and 20% testing data using a fixed random seed for reproducibility.

b. Training data are used for model learning, while test data are reserved for final performance evaluation.

5. Feature Standardization

a. Numerical features are standardized to zero mean and unit variance.

b. This step is necessary because both SVM and K-NN are sensitive to feature scale, and unscaled features could dominate distance computations or margin estimation.

6. Final Preprocessed Dataset

The final dataset remains at 836 records and is fully numeric and standardized, making it suitable for regression modeling [20]. Figure 2 presents an overview of the preprocessing results.

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

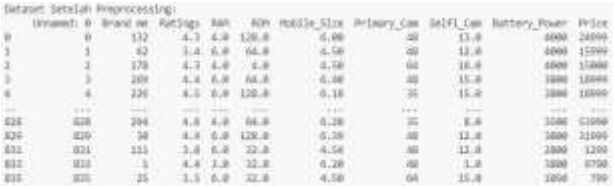


Figure 2. *Dataset Preprocessing Result*

C.Training Data and Test Data Division

The training and testing datasets resulting from the 80:20 split are shown in Figures 3 and 4. The division ensures that model evaluation is performed on unseen data, providing an initial estimate of generalization performance. However, it is acknowledged that a single train–test split may introduce sampling bias; therefore, complementary evaluation using cross-validation is later employed to assess model stability across different data partitions.

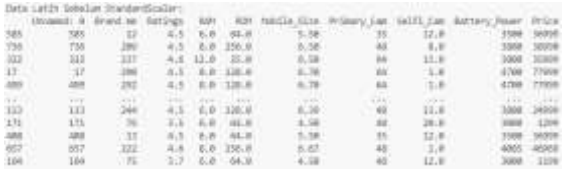


Figure 3. *Training Data*

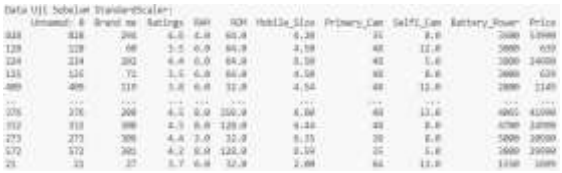


Figure 4. *Test Data*

D.Machine Learning Model Development

The development of the machine learning model in this study was carried out using two main algorithms, namely Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN). The process of building this model consists of several main stages, starting from feature selection to model performance evaluation.

1. The features used in the model include technical specifications that are theoretically and empirically associated with smartphone pricing :

a. Battery Capacity (mAh)

b. Internal Memory (GB)

c. RAM (GB)

d. Main Camera (MP)

e.Secondary Camera (MP)

f.Processor Speed (GHz)

g. Screen Size (Inch)

h. Weight (Gram)

i. 4G and 5G Network Support (Binary: 0/1)

j. Smartphone Brand (Encoded) The target variable to be predicted is the price of the smartphone ("price") in the dataset.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

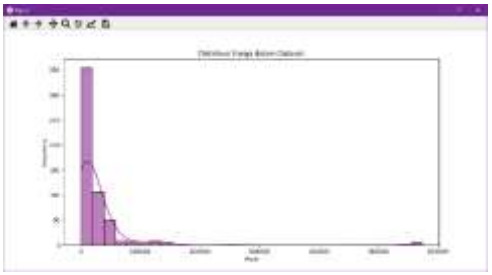


Figure 5. *Price Feature Distribution in Dataset*

The target variable is the smartphone price. The price distribution (Figure 5) is right-skewed, with most observations concentrated in the low-price range. This skewness implies that model performance may vary across price segments, particularly for high-priced devices with fewer training examples [21].

2. Machine Learning Model Development Two machine learning algorithms are used in this study to predict smartphone prices :

a. Support Vector Machine (SVM)

1) A linear kernel is used to model the relationship between features and prices.

2) The model is trained using standardized training data.

3) Predictions are generated on standardized test data.

The prediction results using the SVM model are as follows:

1) Original Price Distribution vs. SVM Prediction

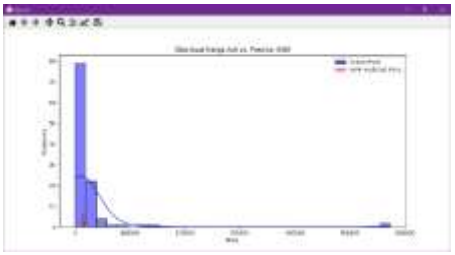


Figure 6. *Original Price Distribution vs. SVM Predictions*

2) SVM Residual Plot

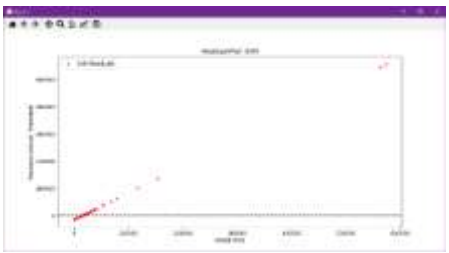
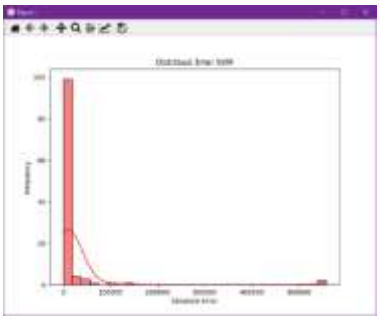


Figure 7. *SVM Residual Plot*

3) SVM Error Distribution



ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

Figure 8. *SVM Error Distribution*

The distribution of SVR predictions (Figure 6) appears less dispersed than the original price distribution, suggesting a tendency toward underestimating extreme values. The residual plot (Figure 7) indicates increasing prediction errors for higher-priced smartphones, implying limited model flexibility. The error distribution (Figure 8) shows that while many predictions have small errors, a non-negligible number of large errors occur, particularly in the upper price range. This pattern suggests potential bias toward the mean and limited capacity of a linear model to capture complex pricing patterns.

b.K-Nearest Neighbors (K-NN)

1) The model uses k = 5 nearest neighbors to estimate smartphone prices.

2) Price predictions are based on local similarity in the standardized feature space.

The prediction results using the SVM model are as follows :

1) Original Price Distribution vs. K-NN Prediction

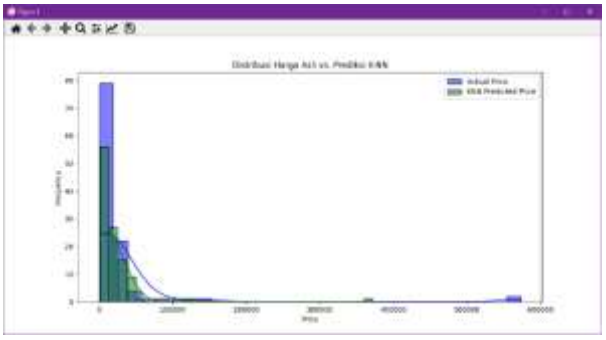


Figure 9. *Original Price Distribution vs. K-NN Prediction*

2) Residual Plot: K-NN

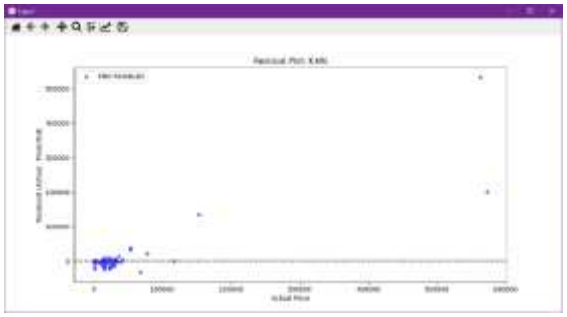


Figure 10. *Residual Plot: K-NN*

3) K-NN Error Distribution

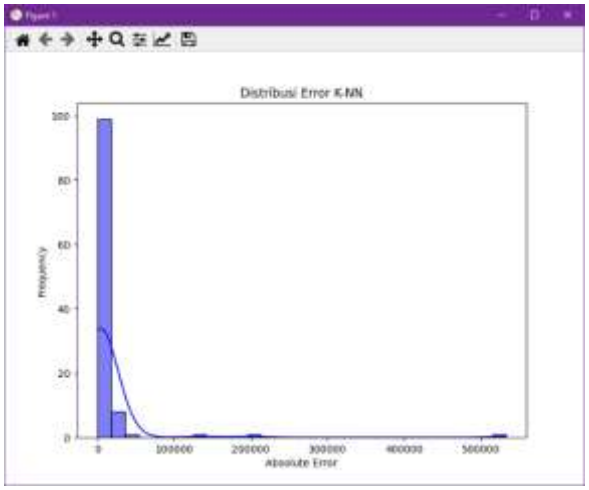


Figure 11. *K-NN Error Distribution*

The predicted price distribution of K-NN (Figure 9) follows the original distribution more closely than SVM, indicating better

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

local pattern capture. The residual plot (Figure 10) shows that most errors are concentrated around zero, although some large deviations occur. The error distribution (Figure 11) confirms that K-NN generally produces smaller errors but remains susceptible to occasional large prediction deviations, especially for smartphones with rare or extreme feature combinations [22].

3. Model Validation with 10-Fold Cross Validation

Cross-validation is conducted to assess model stability across different data partitions :

a. SVM (10-Fold CV R^2 = 0.098)

1) The positive R^2 value indicates that the model captures limited but consistent variance in price across folds.

2) Although predictive accuracy is low, the relatively stable performance suggests lower variance across different samples.

3) The results imply that further optimization, such as kernel selection and hyperparameter tuning, may improve model expressiveness.

b.K-NN (10-Fold CV R^2 = -0.659)

1) The negative R^2 value indicates that the model performs worse than a naive baseline predictor in several validation folds.

2) This outcome reflects high variance and limited generalization, consistent with overfitting behavior.

3) The discrepancy between hold-out test performance and cross-validation results suggests that K-NN is sensitive to data partitioning and local sample density.

E.Model Evaluation

Model evaluation is carried out using three main metrics, namely: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) Score to measure model performance on test data.

1. Support Vector Machine (SVM)

a.MAE = 23,321.39

b.MSE = 5.96×10^9

c.R^2 Score = -0.013

2.K-Nearest Neighbors (K-NN)

a.MAE = 12,809.63

b.MSE = 3.16×10^9

c.R^2 Score = 0.461

Based on the evaluation metrics on the test set, K-NN achieves lower MAE and MSE and a higher R^2 score than SVM, indicating better fit to the observed test data. However, these results should be interpreted cautiously, as cross-validation reveals that K-NN exhibits poor generalization stability. Conversely, SVM demonstrates weaker predictive accuracy but more stable performance across folds. This contrast highlights a trade-off between apparent accuracy on a single split and robustness under repeated sampling [23].

F.Prediction Results

The prediction results are displayed in a table that includes the actual price and the prices predicted by the SVM and K-NN models can be seen in table 1.

Brand	Actual Price	SVM Predicted Price	KNN Predicted Price
Samsung Galaxy S8 Plus (Midnight Black, 64)	53990	15873.46594	21071
I Kall K 15	639	14441.09404	877
OPPO F11 Pro (Aurora Green, 64)	24990	15651.34774	18813.4
I Kall K27 New	639	14482.6259	997
Karbonn k24+	1149	14630.36643	1013.6
...	...	...	...

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

OPPO Reno 10x Zoom (Ocean Green, 256)	41990	16746.69222	39393.6
Vivo V17 (Glacier ice white, 128)	24990	16469.97145	39095.4
Vivo U10 (Thunder Black, 32)	10990	15755.50143	11106.6
Vivo NEX (Black, 128)	29990	16106.11591	24797.2
Blacear I7 Trio	1699	14007.34053	1777.2

Table 1. *Prediction Results*

The comparison between predicted and actual prices shows that SVM tends to underestimate smartphone prices, particularly in the higher price range, while K-NN provides closer approximations for low- to mid-range prices. Nonetheless, both models exhibit larger absolute errors for premium devices, reflecting the limited representation of high-priced smartphones in the dataset. This observation suggests that prediction reliability is not uniform across all price segments [24].

G.Discussion

The following are the results of the evaluation of the two models :

Model	MAE	MSE	R^2 Score
SVM	23,321.39	5.96×10^9	-0.013
K-NN	12,809.63	3.16×10^9	0.461

Table 2. *Model Evaluation Results*

Based on the prediction results, SVM tends to produce lower predicted prices compared to the actual prices, with some cases showing significant differences. Meanwhile, K-NN is closer to the actual prices, especially for smaller prices, but still shows some errors. For example, here are some prediction results from both models:

Brand	Actual Price	SVM Predicted Price	K-NN Predicted Price
Samsung Galaxy S8 Plus	53,990	15,873.46	21,071
I Kall K 15	639	14,441.09	877
OPPO F11 Pro	24,990	15,651.35	18,813.4
OPPO Reno 10x Zoom	41,990	16,746.69	39,393.6
Vivo V17	24,990	16,469.97	39,095.4

Table 3. *Sample Prediction Results*

The comparative analysis demonstrates that K-NN provides better predictive accuracy on the hold-out test set, whereas SVM offers greater stability under cross-validation. The negative cross-validated R^2 of K-NN indicates substantial variance and susceptibility to overfitting, likely due to the instance-based nature of the algorithm and sensitivity to local data density.

The right-skewed price distribution contributes to performance disparities, as both models are biased toward the dominant low-price segment. SVM's tendency to underestimate high prices reflects the limited expressiveness of a linear kernel in capturing complex, non-linear relationships between technical specifications and market prices. However, this interpretation remains a hypothesis, as non-linear kernels were not empirically evaluated in the current configuration [25].

Overall, the findings suggest that model selection should be guided by application context. If the primary objective is minimizing prediction error on known data, K-NN may be preferable. However, for applications requiring more stable generalization to new data, SVM offers a more conservative choice, albeit at the cost of reduced accuracy. Future work should incorporate non-linear kernels, systematic hyperparameter tuning, and stratified error analysis across price ranges and brands to enhance robustness and interpretability.

Although advanced regression models such as Gradient Boosting and ensemble-based methods have demonstrated superior performance in prior studies [6], this study intentionally focuses on SVM and K-NN as baseline models to analyze fundamental differences between margin-based and instance-based learning in smartphone price prediction [26]. The selection of these two algorithms is theoretically justified, as SVM constructs a model based on the principle of margin maximization, whereas K-NN predicts target values based on the proximity and similarity of data points within the feature space. These differences in their underlying mechanisms enable a comparative analysis of distinct modeling approaches for smartphone price prediction. Thus, the use of SVM and K-NN as baseline models is intended not only to compare prediction results but also to understand the characteristics, strengths, and limitations of each algorithm as a foundation for developing more complex models in future research.

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

CONCLUSION

This study compares the performance of two machine learning algorithms, Support Vector Regression (SVM) and K-Nearest Neighbors (K-NN), in predicting smartphone prices based on their technical specifications. Based on the evaluation results, the following conclusions can be drawn :

1.K-NN shows lower error on the test set, but the comparison remains methodologically limited.

a.K-NN achieves lower MAE (12,809.63) and MSE (3.16×10^9) compared to SVR, which records an MAE of 23,321.39 and an MSE of 5.96×10^9 on the hold-out test set.

b.The R^2 Score of K-NN (0.461) indicates moderate explanatory power on the testing data, while SVR yields a negative R^2 (−0.013), suggesting poor fit under the current configuration.

c.However, these results should be interpreted cautiously, as the models were evaluated without systematic hyperparameter optimization and kernel exploration, which limits the validity of a definitive performance comparison.

2. Generalization performance indicates that both models have limitations in capturing unseen data patterns.

a. The 10-Fold Cross Validation results show a negative R^2 for K-NN (−0.659), indicating poor generalization and sensitivity to data partitioning.

b. Although SVM achieves a slightly positive cross-validation R^2 (0.098), this value remains low and does not provide strong evidence of reliable generalization.

c. Therefore, the notion of "model stability" cannot be conclusively claimed for either method based solely on R^2, and broader generalization indicators or repeated validation schemes are required for more robust conclusions.

3. Data characteristics and preprocessing choices influence model behavior and error patterns.

a. The dataset exhibits a right-skewed price distribution, with a dominance of lower-priced smartphones and relatively few high-priced samples, which introduces bias in learning price patterns.

b. Under such distributional imbalance, K-NN tends to adapt more flexibly to local patterns but is prone to overfitting, while SVR with a linear kernel lacks sufficient expressiveness to model non-linear relationships.

c. These findings indicate that model performance is not only determined by the algorithm choice but also by the interaction between data distribution, feature representation, and preprocessing strategy.

4. The current findings highlight important methodological limitations and practical considerations for real-world deployment.

a. The use of a linear kernel in SVM and a fixed k=5 in K-NN restricts model expressiveness, rendering the present comparison preliminary rather than conclusive.

b. The relatively small dataset size and the encoding strategy for categorical variables limit the robustness of the learned patterns and may reduce transferability to newly released smartphone models.

c. From a deployment perspective, K-NN may face scalability and computational constraints in real-time pricing systems, while SVR requires careful kernel and parameter tuning to ensure robustness across evolving market conditions. Therefore, future work should prioritize systematic hyperparameter optimization, stronger baseline comparisons, and evaluation under dynamic market scenarios.

ACKNOWLEDGEMENTS

author expresses gratitude to all those who supported the preparation of this article. This research is intended to contribute to the development of machine learning methods for predicting mobile phone prices based on technical specifications. Artificial Intelligence (AI) was used as an aid for language editing; all outputs were reviewed and verified by the author, who assumes full responsibility for the content and originality of the article.

References

1. R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," Journal of Big Data, vol. 7, Art. no. 52, 2020, doi: 10.1186/s40537-020-00327-4.
2. V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. Springer, 2019.
3. T. Zhang, X. Li, and J. Wang, "Price estimation of smartphones using the k-nearest neighbors algorithm," IEEE Transactions on Consumer Electronics, vol. 67, no. 4, pp. 348–359, 2021.
4. F. Gao, H. Sun, and Y. Liu, "Comparative analysis of machine learning models for housing price prediction," Expert Systems with Applications, vol. 189, Art. no. 116032, 2022, doi: 10.1016/j.eswa.2021.116032.
5. R. Li, X. Deng, and H. Xu, "Predicting vehicle prices using hybrid machine learning models," Applied Intelligence,

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 28 No. 1 (2027): January
DOI: 10.21070/ijins.v28i1.2457

vol. 53, no. 1, pp. 287–303, 2023, doi: 10.1007/s10489-022-03615-7.
6. D. Kim, S. Park, and J. Lee, "Evaluating machine learning algorithms for real estate price prediction: SVM vs. K-NN," Computational Economics, vol. 60, no. 2, pp. 219–234, 2022.
7. M. Lee, K. Han, and B. Cho, "Non-linear price prediction using K-NN and deep learning hybrid models," Neural Computing and Applications, vol. 35, no. 5, pp. 8891–8904, 2023.
8. H. A. Abu Alfeilat et al., "Effects of distance measure choice on k-nearest neighbor classifier performance: A review," Big Data, vol. 7, no. 4, pp. 221–248, 2019, doi: 10.1089/big.2018.0175.
9. F. Al-Turjman, A. A. Hussain, S. Alturjman, and C. Altrjman, "Vehicle price classification and prediction using machine learning in the IoT smart manufacturing era," Sustainability, vol. 14, no. 15, Art. no. 9147, 2022, doi: 10.3390/su14159147.
10. N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous datasets," SN Applied Sciences, vol. 1, no. 12, Art. no. 1356, 2019, doi: 10.1007/s42452-019-1356-9.
11. A. Altan and S. Karasu, "The effect of kernel functions in support vector machines on forecasting performance of financial time series," Journal of Cognitive Systems, vol. 4, no. 1, pp. 17–21, 2019.
12. J. Baumann, L. Zavolokina, and G. Schwabe, "Dealers of peaches and lemons: How can used car dealers use trusted car data to create value?," in Proc. 54th Hawaii International Conference on System Sciences, 2021, pp. 5409–5418, doi: 10.24251/hicss.2021.658.
13. C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, no. 3, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.
14. M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine learning for industrial applications: A comprehensive literature review," Expert Systems with Applications, vol. 175, Art. no. 114820, 2021, doi: 10.1016/j.eswa.2021.114820.
15. M. Bilen, "Predicting used car prices with heuristic algorithms and creating a new dataset," Journal of Multidisciplinary Developments, vol. 6, no. 1, pp. 1–12, 2021.
16. S. Y. Chaganti, I. Nanda, K. R. Pandi, T. G. Prudhvith, and N. Kumar, "Image classification using SVM and CNN," in Proc. 2020 ICCSEA, 2020, doi: 10.1109/ICCSEA49143.2020.9132851.
17. A. Dutulescu et al., "Car price quotes driven by data: Comprehensive predictions grounded in deep learning techniques," Electronics, vol. 12, no. 14, Art. no. 3083, 2023, doi: 10.3390/electronics12143083.
18. M. Andan and D. Ghosh, "Pre-owned car price prediction using machine learning techniques," Journal of Decision Analytics and Intelligent Computing, vol. 3, no. 1, pp. 167–184, 2023, doi: 10.31181/jdaic10008102023n.
19. C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies," Remote Sensing, vol. 11, no. 2, Art. no. 185, 2019, doi: 10.3390/rs11020185.
20. N. A. Sinaga, B. H. Hayadi, and Z. Situmorang, "Comparison of Naïve Bayes, K-NN, and SVM accuracy in employee acceptance prediction," Tekinkom, vol. 5, no. 1, pp. 27–34, 2022, doi: 10.37600/tekinkom.v5i1.446.
21. P. Wang, X. Yang, and L. Chen, "Regression-based price prediction for electronic devices," Journal of Big Data, vol. 8, no. 1, Art. no. 75, 2021, doi: 10.1186/s40537-021-00471-3.
22. Y. Zhou, C. Lin, and D. Wu, "A review of AI-based pricing strategies in consumer goods," Artificial Intelligence Review, vol. 58, no. 3, pp. 1451–1473, 2024.
23. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
24. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
25. M. Kuhn and K. Johnson, Applied Predictive Modeling. Springer, 2013.
26. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. O'Reilly Media, 2019.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).