

Prediction of Dirty Cryptocurrency Using Machine Learning Models

Lina He^{1,a,*}

¹*Mathematics, Statistics and Actuarial Science, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK*

a. lh22684.essex.ac.uk

**corresponding author*

Abstract: Predicting the value of cryptocurrencies is essential due to their significant impact on the financial market and investment strategies. This paper evaluates the performance of three machine learning models: K-Nearest Neighbor (KNN), Linear Regression, and Random Forest Regression in predicting the value of various cryptocurrencies (Bitcoin, Ethereum, Bitcoin Cash, Litecoin, and Binance Coin (BNB)). Important metrics such as R-squared (R^2), accuracy, precision, recall, F1 score, Mean Squared Error (MSE), and Mean Absolute Error (MAE) are utilized alongside graphical comparisons of actual and predicted price categories. The analysis reveals that Random Forest Regression performs best for Bitcoin, while KNN is more suitable for Ethereum, BNB, and Litecoin. Linear Regression models tend to deviate more from actual prices. Both KNN and Random Forest Regression demonstrate a significantly closer relationship to real price categories. This study underscores the importance of model selection, highlighting the superior accuracy of KNN and random forest in predicting cryptocurrency prices. Future work could explore the integration of more advanced models and the impact of additional features on prediction accuracy.

Keywords: Cryptocurrency, K-Nearest Neighbor, Linear Regression, Random Forest Regression.

1. Introduction

The rapid rise of cryptocurrencies has revolutionized the financial landscape, creating new opportunities and difficulties for experts and investors alike. However, several cryptocurrencies pose significant environmental risks due to their large carbon footprint and high energy consumption. Mining cryptocurrencies such as Bitcoin and Ethereum requires significant processing power, resulting in significant energy consumption and environmental damage. Accurate price prediction models are essential to maximize trading strategies and minimize meaningless trading, both of which can reduce the negative environmental impact of HFT. Reliable forecasting models help investors, financial analysts, and environmental advocates to make better judgments to balance the sustainability of environmental and financial goals [1].

To mitigate the negative environmental impact of cryptocurrencies, accurate price estimates are necessary. Machine learning models have been widely used in the field of financial forecasting because they can deal with complex nonlinear patterns in huge data sets. We evaluate the effectiveness of three machine learning models: K-Nearest Neighbor (KNN), Linear Regression, and

Random Forest Regression to predict the prices of five important cryptocurrencies: BNB, Litecoin, Ethereum, Bitcoin Cash, and Bitcoin. These cryptocurrencies were chosen because of their wide application, large market share, and varying degrees of environmental impact. KNN is known for its ease of use and efficient discovery of patterns in local data. Linear Regression is a simple way to model the correlation between variables. Random Forest Regression, an ensemble learning technique, combines multiple decision trees for high robustness and accuracy [2].

Which machine learning model has the best forecasting accuracy for the prices of well-known cryptocurrencies? This study aims to provide an answer to this research question. We examined historical pricing data from July 1, 2017, to July 1, 2022, using daily closing prices. The performance of each model was evaluated using metrics such as R-squared (R-squared), accuracy, precision, recall, F1 score, mean squared error (MSE), and Mean Absolute error (MAE). The predictive power of the model can be seen by graphical comparison of actual and predicted prices. Our results support environmentally friendly trading methods, reduce the impact of cryptocurrencies with significant environmental impact, and help investors and data scientists select appropriate machine learning models for Bitcoin price prediction [3, 4].

2. Methodology

2.1. K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) is an easy-to-use, example-based learning method that can be applied to regression and classification problems. Its function is to determine which k data points are closest to the query location, generate predictions based on those neighbors, and average the values of those nearest neighbors for a local view. Because of this approach, KNN is particularly useful in identifying local patterns and short-term trends. The algorithm is easy to apply and does not require any assumptions about how the data is distributed. However, it is sensitive to extraneous features and data sizes, and it is computationally expensive for large data sets.

2.2. Linear Regression

Linear Regression is a basic statistical method that fits linear equations into observational data to simulate the relationship between a dependent variable and one or more independent variables. By establishing linear relationships across datasets, it provides a global perspective that helps to understand overall trends and relationships. The method is easy to understand and interpret, and it is also computationally efficient. Nevertheless, it assumes that the relationship between variables is linear, which may not adequately capture more complex patterns in the data.

2.3. Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees and combines their outputs to enhance accuracy and reduce overfitting. The technique provides a balanced approach by combining numerous decision trees to capture complex non-linear links and interactions in data. The advantages of random forests include efficient handling of large datasets and high-dimensional spaces, resilience to overfitting, and the ability to provide recognition of the importance of individual features. It has lower interpretability and higher computational requirements compared to simple models [5].

3. Data

The analysis uses historical price data from January 2017 to December 2022, with daily closing prices for five major cryptocurrencies: Bitcoin, Ethereum, Bitcoin Cash, Litecoin, and BNB. The cryptocurrencies are selected due to their significant market presence, diverse use cases, and varying levels of environmental impact [6]. The data is used to evaluate the performance of the three machine learning models KNN, Linear Regression, and Random Forest Regression in predicting the prices of these cryptocurrencies. Important metrics such as R^2 , accuracy, precision, recall, F1 score, MSE, and MAE are utilized to assess the models' effectiveness [7-9]. All the datasets have been downloaded from the investing, the Sample Index is used to represent each observation in the datasets. This approach is particularly useful when dealing with large datasets, as using dates directly on the x-axis can make the chart too dense and difficult to read. The Sample Index allows for a clearer comparison between the actual and predicted values at various points in the datasets, ensuring that the visualization remains clear and comprehensible. This helps in effectively displaying the model's performance and the differences between actual and predicted values [10].

4. Results

4.1. The Prediction of Bitcoin

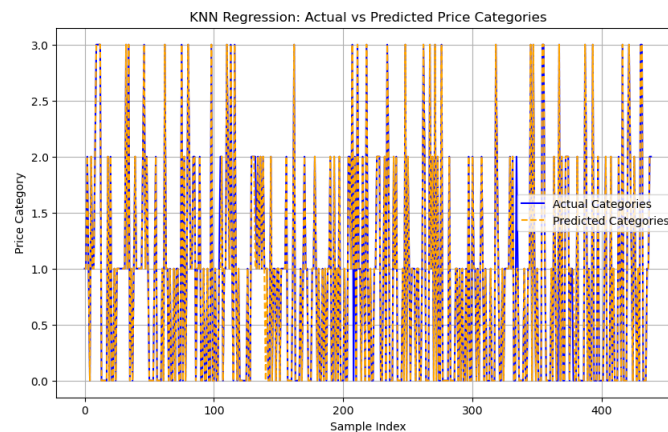


Figure 1: Actual vs Predicted Bitcoin Prices Using KNN

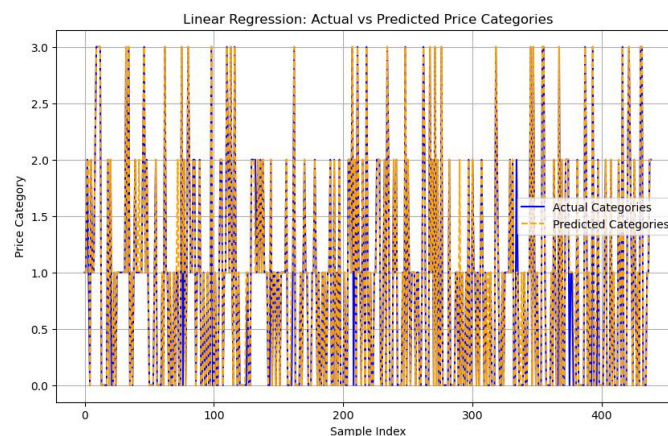


Figure 2: Actual vs Predicted Bitcoin Prices Using Linear Regression

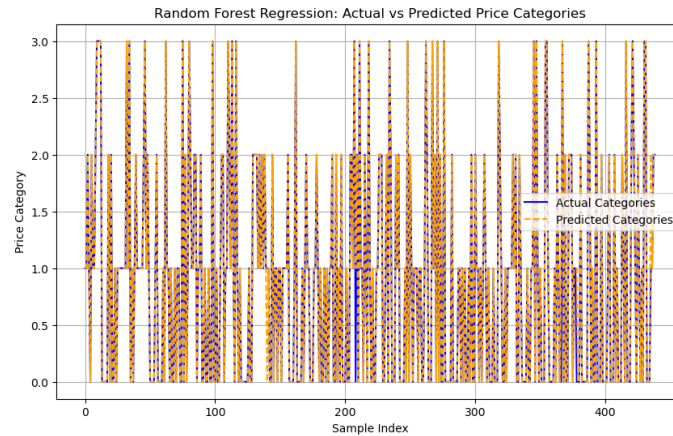


Figure 3: Actual vs Predicted Bitcoin Prices Using Random Forest Regression

The Figures given provides a comparison of the expected and actual price categories (blue and orange dashed lines) for Bitcoin across the three models. Figure 2 do a poor job of capturing the complexity of Bitcoin's price movements, as evidenced by frequent and large discrepancies between actual and predicted classes. The Figure 3 shows that the real and predicted categories are closer together and the variance is smaller, indicating higher prediction accuracy. The high alignment on the Figure 1 indicates high prediction accuracy with few outliers. Consistent with numerical metrics, graphical comparisons show that Random Forest Regression and KNN are the best models for predicting Bitcoin's value, while Linear Regression underperforms.

Table 1: Performance Metrics for Bitcoin Price Prediction

Bitcoin	KNN	Linear Regression	Random Forest Regression
R ²	0.9830	0.9268	0.9888
accuracy	0.9841	0.9544	0.9886
precision	0.9845	0.9577	0.9887
recall	0.9841	0.9544	0.9886
F1 Score	0.9841	0.9549	0.9886
MSE	0.0157	0.0677	0.0104
MAE	0.0282	0.2077	0.0170

The Table 1 shows the key performance indicators for predicting the price of Bitcoin using Linear, Random Forest, and KNN regression. The R² value was the largest (0.9888), and the fitting effect of Random Forest Regression was the strongest, followed by Linear Regression (0.9268) and KNN regression (0.9830). The best accuracy was achieved by Random Forest Regression (0.9886), which was followed by Linear Regression (0.9544) and KNN (0.9841). There were also greater accuracy, recall, and F1 score with KNN (precision: 0.9845, recall: 0.9841, F1 Score: 0.9841) and Linear Regression (precision: 0.9577, recall: 0.9544, F1 Score) than with Random Forest Regression (precision: 0.9887, recall: 0.9886, F1 Score: 0.9886). The errors of Random Forest Regression and KNN are smaller than Linear Regression (MSE is 0.0677, MAE is 0.2077), MSE is 0.0104 and 0.0157, and MAEs is 0.0169 and 0.0282, respectively. These metrics illustrate the superior performance of Random Forest Regression in predicting the price of Bitcoin.

4.2. The Prediction of Ethereum

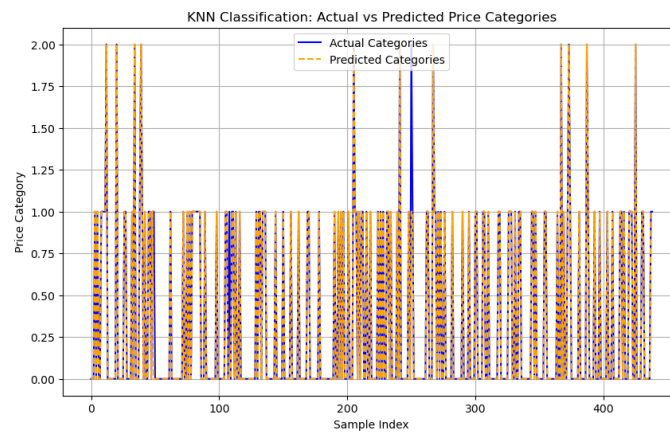


Figure 4: Actual vs Predicted Ethereum Prices Using KNN

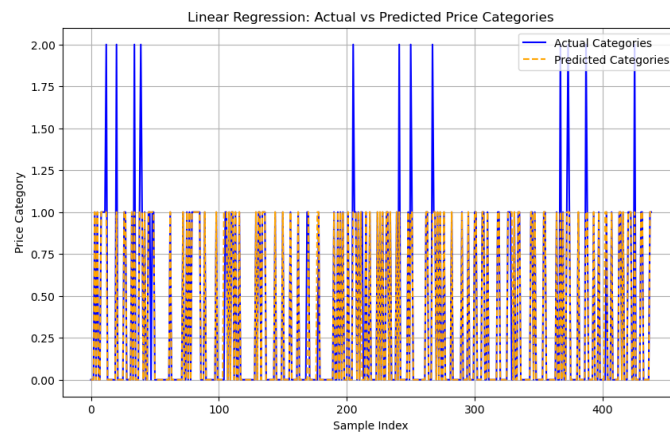


Figure 5: Actual vs Predicted Ethereum Prices Using Linear Regression

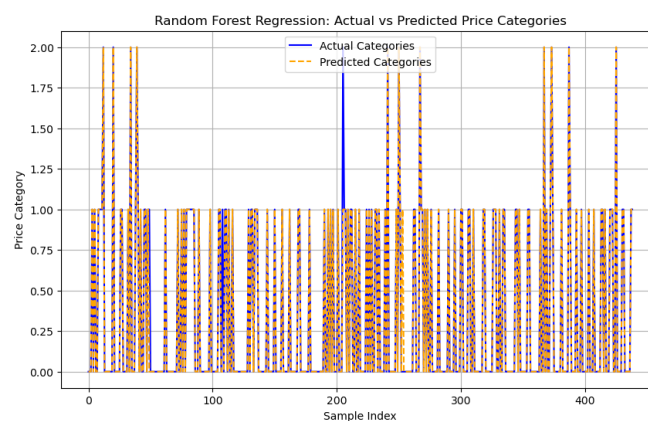


Figure 6: Actual vs Predicted Ethereum Prices Using Random Forest Regression

The above Figures (figure 4, figure 5, and figure 6) compares the three models' predictions for the Ethereum price class (orange dashed line) with the actual price class (blue line). The Linear Regression graph does not well reflect the complexity of Ethereum's price movements, and it often deviates significantly from the expected category. The Random Forest Regression plot shows that the

real and predicted categories are closer together and the variance is smaller, indicating higher prediction accuracy. The high alignment on the KNN Regression plot indicates high prediction accuracy with few outliers. The numerical metrics were confirmed by visual comparisons, which showed that KNN and Random Forest Regression were the best models for predicting Ethereum's value, while Linear Regression performed poorly.

Table 2: Performance Metrics for Ethereum Price Prediction

Ethereum	KNN	Linear Regression	Random Forest Regression
R ²	0.9750	0.9567	0.9712
accuracy	0.9932	0.9337	0.9909
precision	0.9932	0.9337	0.9909
recall	0.9932	0.9567	0.9909
F1 Score	0.9931	0.9443	0.9908
MSE	0.0068	0.0783	0.0079
MAE	0.0068	0.2307	0.0119

The Table 2 shows important performance metrics for predicting Ethereum prices using KNN, Linear Regression, and Random Forest Regression. The R² value of KNN was the largest (0.9750), followed by Random Forest Regression (0.9712), and the R² value of Linear Regression was smaller (0.9567), which was a good fit to the data. KNN has the greatest accuracy (0.9932), followed by Random Forest Regression (0.9909) and Linear Regression (0.9337). In comparison to Random Forest Regression (precision: 0.9909, recall: 0.9909, F1 Score: 0.9908) and Linear Regression (precision: 0.9337, recall: 0.9567, F1 Score: 0.9443), KNN has greater precision, recall, and F1 Score (precision: 0.9932, recall: 0.9932, F1 Score: 0.9931). The errors of KNN and Random Forest Regression were lower than Linear Regression (MSE: 0.0783, MAE: 0.2307), and MSE and MAEs were 0.0068 and 0.0079, 0.0068 and 0.0119, respectively. These indicators reflect the predictive effect of KNN and Random Forest Regression on Ethereum prices.

4.3. The Prediction of Bitcoin Cash

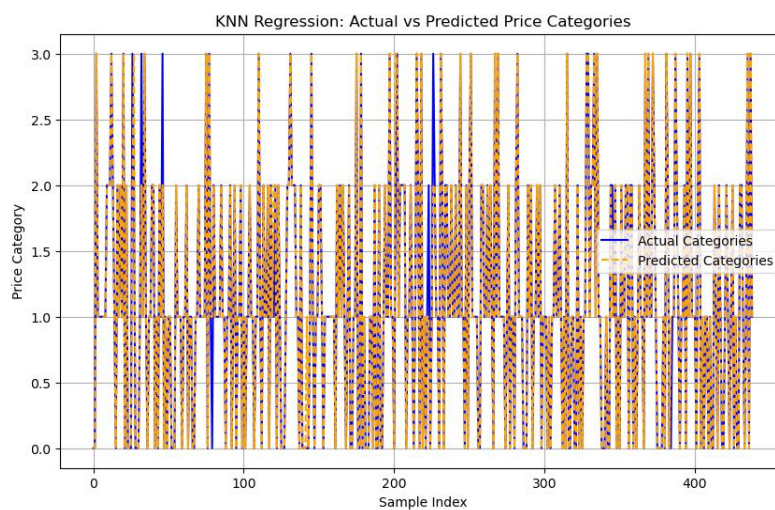


Figure 7: Actual vs Predicted Bitcoin Cash Prices Using KNN

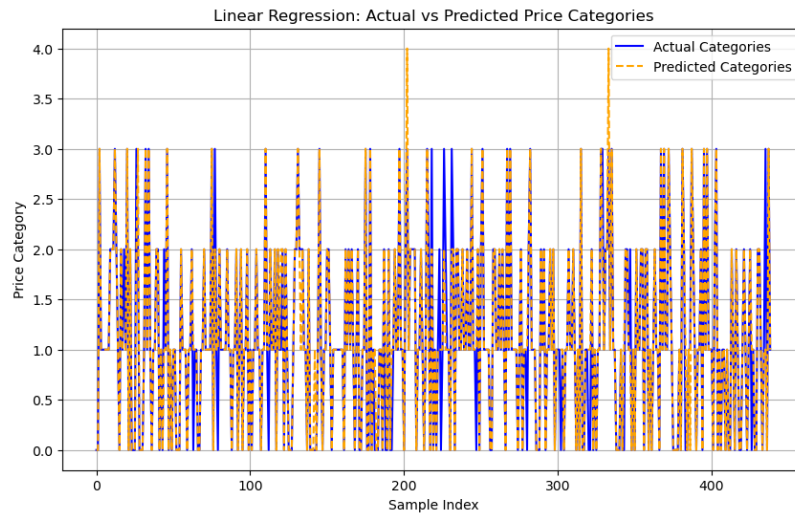


Figure 8: Actual vs Predicted Bitcoin Cash Prices Using Linear Regression

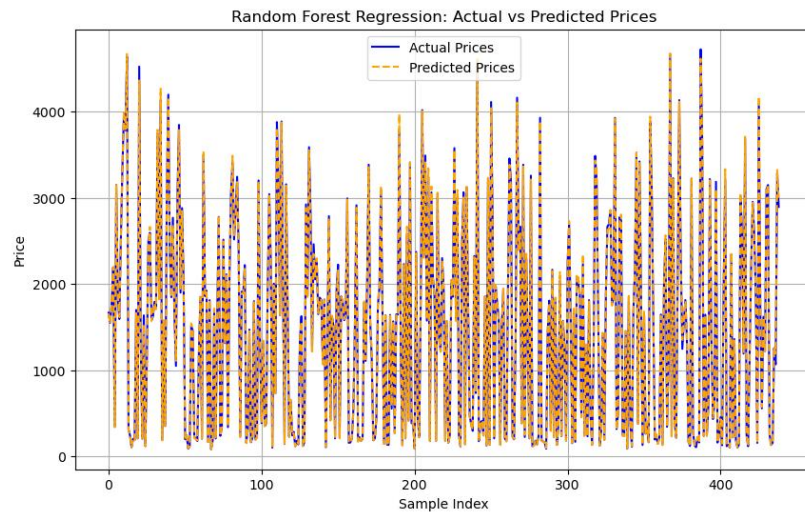


Figure 9: Actual vs Predicted Bitcoin Cash Prices Using Random Forest Regression

The Figures provided provide a comparison of Bitcoin cash between the expected price and actual price categories (blue and orange dashed lines) for the three models (see Figure 7, figure 8, and figure 9). It is evident from the Linear Regression plots that the model is unable to adequately capture the complexity of Bitcoin cash price volatility, with frequent and significant deviations from actual and expected categories. Like The Figure 7 and 9, include significant deviations as well, indicating potential limits to its forecast ability. The Figure 7, on the other hand, shows a good degree of predictive accuracy with minimum deviations, showing a high degree of alignment between actual and projected categories. The numerical metrics are corroborated by the visual comparison, which shows that KNN is the best model for predicting the price of Bitcoin Cash, with Linear Regression and Random Forest Regression doing less well.

Table 3: Performance Metrics for Bitcoin Cash Price Prediction

Bitcoin cash	KNN	Linear Regression	Random Forest Regression
R ²	0.9780	0.8950	0.8950
accuracy	0.9727	0.9203	0.9203
precision	0.9728	0.9297	0.9297
recall	0.9727	0.9203	0.9203
F1 Score	0.9726	0.9217	0.9217
MSE	0.0187	0.0891	0.0891
MAE	0.0296	0.2481	0.2481

The table 3 shows the important performance indicators for predicting the price of Bitcoin Cash using KNN, Linear Regression, and Random Forest Regression. The R² values of Linear Regression and Random Forest Regression are both small (0.8950), while KNN has the highest R² value (0.9780), which fits the data well. The same pattern can be seen in accuracy, where KNN beats both Linear Regression and Random Forest Regression with a score of 0.9727, considerably surpassing both at 0.9203. In comparison to Linear Regression and Random Forest Regression (precision: 0.9297, recall: 0.9203, F1 score: 0.9217), KNN has comparable higher precision, recall, and F1 Scores (precision: 0.9728, recall: 0.9727, F1 Score: 0.9726). KNN has an MSE of 0.0187 and a MAE of 0.0296, which is also the lowest error. Linear Regression and Random Forest have larger errors (MSE of 0.0891 and MAE of 0.2481). These indicators show that the KNN model is better than the other two models in predicting the price of Bitcoin Cash.

4.4. The Prediction of BNB

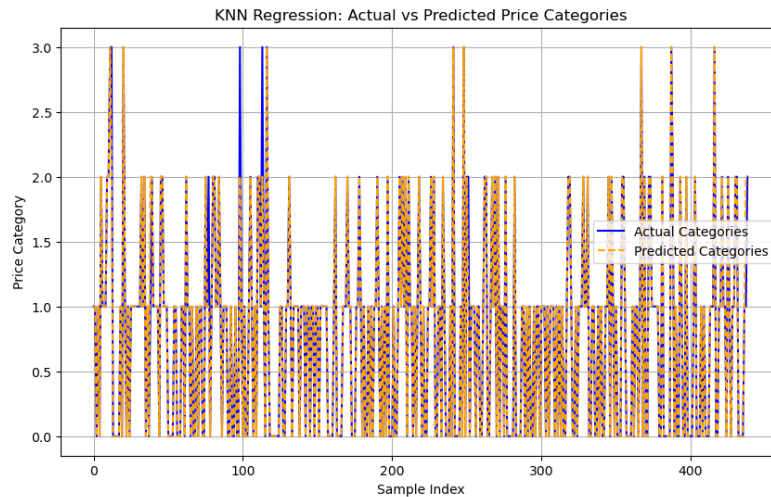


Figure 10: Actual vs Predicted BNB Prices Using KNN

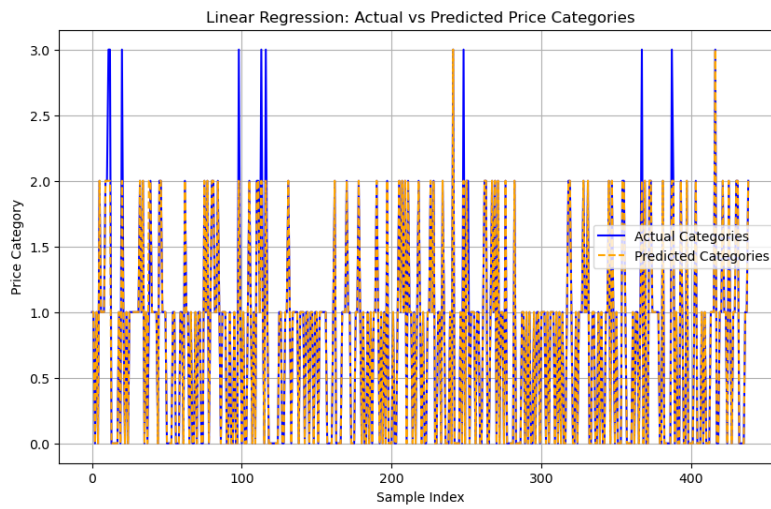


Figure 11: Actual vs Predicted BNB Prices Using Linear Regression

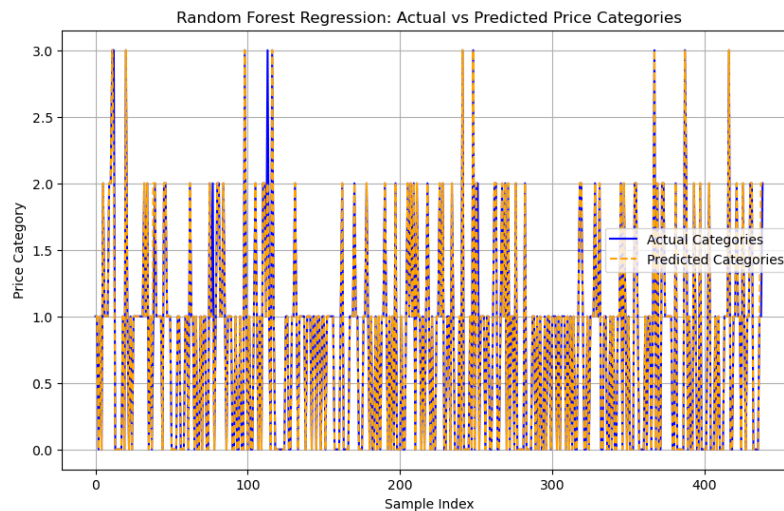


Figure 12: Actual vs Predicted BNB Prices Using Random Forest Regression

For BNB, Figure 11 shows frequent and significant deviations, indicating poor performance. In contrast, Figure 10 and Figure 11 show close alignment between actual and predicted categories, reflecting their strong predictive accuracy and ability to capture data variations effectively. The visual comparison supports the numerical metrics, highlighting KNN and Random Forest as more effective models for predicting BNB prices than Linear Regression (see Figure 12).

Table 4: Performance Metrics for BNB Price Prediction

BNB	KNN	Linear Regression	Random Forest Regression
R ²	0.9826	0.9368	0.9819
accuracy	0.9841	0.9749	0.9863
precision	0.9846	0.9779	0.9866
recall	0.9841	0.9749	0.9863
F1 Score	0.9837	0.9687	0.9862
MSE	0.0106	0.0383	0.0110
MAE	0.0164	0.1365	0.0157

The table 4 shows the performance indicators of KNN, Linear Regression, and Random Forest Regression in predicting BNB prices. Linear Regression (R^2 : 0.9368, accuracy: 0.9749, precision: 0.9779, recall: 0.9749, F1 Score: 0.9687, MSE: 0.0383, MAE: 0.1365), KNN (r^2 : 0.9749, recall: 0.9841, accuracy: 0.9846, recall: 0.9841, F1-score: 0.9837, MSE: 0.0106, MAE: 0.0164) and random forest (r -squared: 0.9819, precision: 0.9863, precision: 0.9866, recall: 0.9863, F1-score: 0.9862, MSE: 0.0109, MAE: 0.0164) were significantly better than Random Forest (R^2 : 0.9819, precision: 0.9866, recall: 0.9863, F1-score: 0.9862, MSE: 0.0109, MAE: 0.0164). These indicators show that Random Forest Regression and KNN have good predictive power for BNB pricing.

4.5. The Prediction of Litecoin

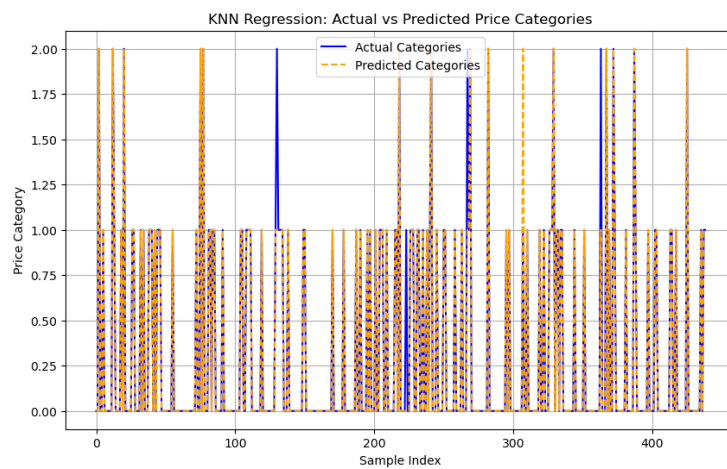


Figure 13: Actual vs Predicted Litecoin Prices Using KNN

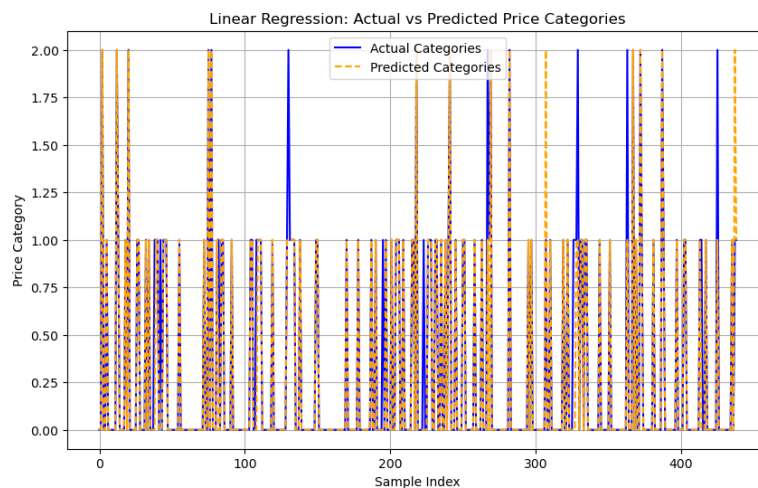


Figure 14: Actual vs Predicted Litecoin Prices Using Linear Regression

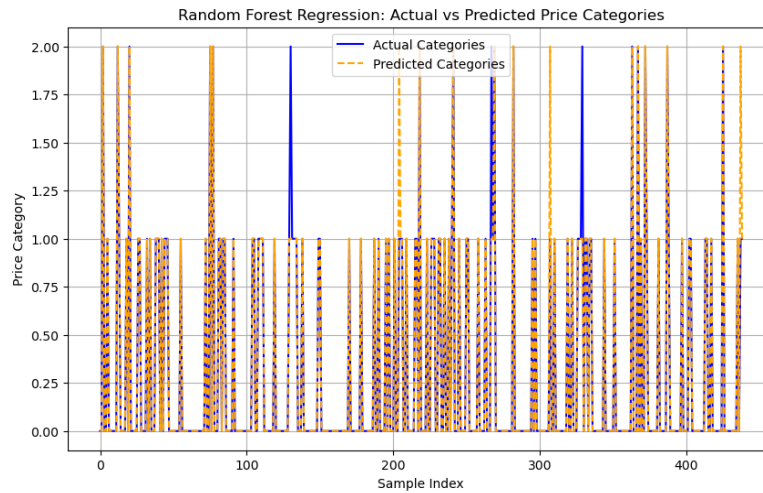


Figure 15: Actual vs Predicted Litecoin Prices Using Random Forest Regression

The Figures (see Figure 13, 14, and 15) above compares the predicted prices of Litecoin with the actual prices. Due to the small variance, KNN demonstrates a strong consistency between actual and predicted values. Additionally, the Random Forest model is tightly packed, indicating its ability to handle data fluctuations effectively. On the other hand, Linear Regression does not capture the non-linear trend in Litecoin's price, which shows frequent deviations. Numerical measurements support the visual study, which shows that KNN and Random Forest perform better than Linear Regression at predicting Litecoin value.

Table 5: Performance Metrics for Litecoin Price Prediction

Litecoin	KNN	Linear Regression	Random Forest Regression
R ²	0.9664	0.8143	0.9618
accuracy	0.9886	0.9613	0.9863
precision	0.9884	0.9600	0.9863
recall	0.9886	0.9613	0.9863
F1 Score	0.9884	0.9599	0.9863
MSE	0.0095	0.0524	0.0108
MAE	0.0173	0.1858	0.0178

The table 5 shows the key performance metrics for each model, including R², precision, recall, F1 score, mean squared error, and mean absolute error. Notably, Random Forest and KNN Regression exhibit high R² values (0.9664 and 0.9618), while Linear Regression exhibits lower R² values (0.8143), indicating a higher robustness of the matching. The accuracy of KNN method was the highest (0.9886), followed by Linear Regression method (0.9613) and Random Forest method (0.9863). While Linear Regression has lower precision, recall, and F1-score (0.9599), KNN and Random Forest have similar scores (0.9884 and 0.9863). Compared to Linear Regression (MSE: 0.0524, MAE: 0.1858), KNN and Random Forest also have smaller errors (MSE: 0.0095 and 0.0108, MAE: 0.0173 and 0.0178). The experimental results show that Random Forest and KNN have superior performance in Litecoin price prediction.

4.6. Summary of Best Model Performance

Table 6: Best Model Performance for Cryptocurrency Price Prediction

Cryptocurrency	Best Model	R ²	accuracy	precision	recall	F1 Score	MSE	MAE
Bitcoin	Random Forest	0.9888	0.9886	0.9887	0.9886	0.9886	0.0104	0.0170
Ethereum	KNN	0.9750	0.9932	0.9932	0.9932	0.9931	0.0068	0.0068
Bitcoin Cash	KNN	0.9780	0.9727	0.9728	0.9727	0.9726	0.0187	0.0296
Litecoin	KNN	0.9664	0.9886	0.9884	0.9886	0.9884	0.0095	0.0173
BNB	KNN	0.9826	0.984	0.9846	0.9841	0.9837	0.0106	0.0164

KNN and Random Forest Regression models are highly effective for predicting the prices of most cryptocurrencies. KNN is particularly strong for Ethereum, Bitcoin Cash, Litecoin, and Binance Coin, while Random Forest Regression excels with Bitcoin. Linear Regression, while simple and computationally efficient, generally underperforms in comparison to the other two models, particularly in capturing the non-linear patterns prevalent in cryptocurrency price data. This analysis suggests that using advanced models like KNN and Random Forest can significantly improve prediction accuracy, aiding investors in making better-informed decisions.

5. Conclusion

Machine learning models vary in their effectiveness in predicting the price of various cryptocurrencies, such as Bitcoin, Ethereum, Bitcoin Cash, Litecoin, and BNB. These models include KNN, Linear Regression, and Random Forest Regression. Of the models tested, Random Forest Regression has the best R² and accuracy, and generally performs better against Bitcoin. For Ethereum, BNB, Litecoin, and Bitcoin Cash, KNN works best, generally exhibiting the highest accuracy and lowest error rate. Graphical analysis confirmed these results, showing that KNN and Random Forest Regression models are closer to the actual pricing categories than Linear Regression models, which often deviate. These findings highlight the importance of using machine learning models that fit the unique properties of Bitcoin data. Investors and data scientists can make more informed decisions by utilizing KNN and Random Forest Regression to generate more accurate price forecasts and potentially reduce the environmental impact of trading dirty cryptocurrencies.

References

- [1] Duan, K., Zhao, Y., Urquhart, A., & Huang, Y. (2023). Do clean and dirty cryptocurrencies connect with financial assets differently? The role of economic policy uncertainty. *Energy Economics*, 127, 107079.
- [2] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, 5(1), 12.
- [3] Gudavalli, H. N., & Kancharla, K. V. R. (2023). Predicting Cryptocurrency Prices with Machine Learning Algorithms: A Comparative Analysis. Retrieved from <https://urn.kb.se/resolve?urn=urn:nbn:se:bth-25077>.
- [4] McNally, S., Roche, J., & Caton, S. (2018). Predicting the price of bitcoin using machine learning. In *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*. IEEE. 339-343.
- [5] MK, M. M., Iswariya, B., Prasad, H., & Sudhakar, T. D. (2024). Load Forecasting Using Random Forest Regression Algorithm in Machine Learning. In *2024 International Conference on Science Technology Engineering and Management (ICSTEM)* IEEE, 1-6.
- [6] Aggarwal, S., & Kumar, N. (2021). Cryptocurrencies. In *Advances in Computers*, 121, 227-266.
- [7] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, 7, e623.

- [8] Junker, M., Hoch, R., & Dengel, A. (1999). *On the evaluation of document analysis components by recall, precision, and accuracy*. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99* (Cat. No. PR00318). IEEE, 713-716.
- [9] Humphrey, A., Kuberski, W., Bialek, J., Perrakis, N., Cools, W., Nuytens, N., ... & Cunha, P. A. C. (2022). *Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth*. *Monthly Notices of the Royal Astronomical Society: Letters*, 517(1), L116-L120.
- [10] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, New York: springer. 2, 1-758.