ANALYSIS OF THE APPLICATION OF HYPERPARAMETER TUNING IN MACHINE LEARNING TO INCREASE THE ACCURACY OF SALES-LEVEL PREDICTION

Sugiyanti[1], Muhammad Haris[2]

[1] Magister of Computer Science, Faculty of Information Technology, Nusa Mandiri University,
[2], Faculty of Information Technology, Nusa Mandiri University

[1,2] Kramat Raya Steet No.18 RW.7 Kwitang Senen, Jakarta Pusat, 10450, Indonesia
[1] 14210253@nusamandiri.ac.id, [2] muhammad.uhs@nusamandiri.ac.id

Abstract

The growth of sales business actors is increasing, so there is a need to predict sales levels for future sales so as not to experience financial losses. This research aims to obtain an accurate sales-level prediction model. The seller will understand the important features that influence the level or value of sales. The method used in this research is the Machine Learning (ML) regression algorithm, which uses hyperparameter tuning. The preprocessing stage in this research is very important to produce better prediction values. This research produced the best algorithm, namely XGBoost, with a Root Mean Squared Error (RMSE) result of 968 and a Mean Absolute Error (MAE) value of 713. These results are better than those of previous research using the same XGBoost algorithm but not optional tuning hyperparameters, namely RMSE of 1052 and MAE of 739.03. So by using optuna tuning hyperparameters you can reduce the error value of the prediction results.

***Keywords** : Predictions, Sales, Hyperparameters, Optuna, XGBoost*

Abstrak

Pertumbuhan pelaku bisnis penjualan semakin meningkat, untuk itu perlu adanya suatu prediksi terhadap tingkat penjualan untuk penjualan masa depan agar tidak mengalami kerugian finansial. Tujuan dari penelitian ini adalah untuk mendapatkan model prediksi tingkat penjualan dengan akurat. Pihak penjual akan memahami fitur-fitur penting yang mempengaruhi terhadap tingkat atau nilai penjualan. Metode yang digunakan dalam peneltiian ini adalah algoritma regression *Machine Learning (ML)* serta menggunakan *hyperparameter tuning*. Tahap *preprocessing* dalam penelitian ini sangat penting untuk dapat menghasilkan nilai prediksi yang lebih baik. Hasil dari penelitian ini menghasilkan algoritma terbaik yaitu XGBoost dengan hasil *Root Mean Squared Error (RMSE)* adalah 968 dan nilai *Mean Absolute Error (MAE)* adalah 713, hasil tersebut lebih baik dari hasil penelitian sebelumnya dengan menggunalan algoritma yang sama yaitu *XGBoost* tetapi tidak menggunakan *hyperparameter tuning optuna* menghasilkan nilai *RMSE* sebesar 1052, dan *MAE* sebesar 739.03. Sehingga dengan menggunakan *hyperparameter tuning optuna* dapat menurunkan nilai *error* hasil prediksi.

Kata Kunci : Prediksi, Penjualan, *Hyperparameter, Optuna, XGBoost*

1. INTRODUCTION

A period of transformation supported by digitalization, information, communication technology, and artificial intelligence has resulted in rapid technological developments [1]. Increasingly sophisticated technology can be helpful in various fields, such as business and economics, resulting in many researchers arguing that increasingly sophisticated technology will usher in a new era [1]. In the field of business and economics, sales activities are undoubtedly inseparable. Sales is an action carried out by a company or retail store to maintain its business so that it continues to grow and obtain the desired profits. Therefore, to maintain maximum profits, an accurate prediction of the level or value of sales needs to be made [2].

Some retail stores still use simple technology, such as Microsoft Excel, to estimate sales, sometimes based only on the leader's wishes and only on expected estimates [1]. Sellers will need data regarding demand for products currently on the market to decide whether to increase or decrease the number of product units. Business people will constantly improve sound quality to satisfy consumers with the products or goods they buy. Business people will also compete in the market for the products or goods they own. If a business competes in the market without considering principles, then the business will be at risk of experiencing financial losses.

With sophisticated technology, it can enable companies to carry out accurate sales level prediction analysis. A business company needs to have the ability to estimate product sales. Accurate predictions or forecasting can help businesses in companies maximize investment, reduce inventory costs, increase sales and profitability, and avoid danger [3]. With accurate predictions about how sales will be in the future, business growth can be adequately managed and will have fatal consequences for sales [3]. Prediction is an action to predict circumstances or conditions for the future by testing past data. Sales prediction is a forecast of future sales based on past sales that occurred [4]. By predicting the level or value of sales, companies can make reliable and appropriate decisions and optimize their business resources. The advantage of business companies in making predictions on product sales data is that they can determine the current level of demand for services or products in the market and future demand for products needed by customers.

Predicting the level or value of sales cannot be done just like that, and experts are needed. That is, they need to pay attention to the features that influence it, such as product type, price, type of shop, location, etc. Therefore, predicting the level or value of sales is influenced by many features. However, these features still need sufficient support in predicting sales, so a prediction model for the level or value of sales is needed accurately based on existing features.

This research will concentrate on forecasting the level or value of sales in a store. Product level or value hypotheses impact sales, such as brand, price, visibility, and so on, while store hypotheses, such as city type, store capacity, location, etc.

There needs to be a detector regarding the level or value of product sales in a shop. One prediction method that is widely used and is considered to be more effective and efficient than others is Machine Learning (ML). Another method that can be used apart from ML is statistical or traditional methods. Still, this method requires a lot of time and is expensive, whereas using a statistical approach focuses on traditional mathematical methods. The difference between using ML and statistical methods is that methods using statistics are more rigid, while using ML can adapt to the problem to be detected. ML algorithms are currently widely used by large companies worldwide for predictions. A collection of statistical methods known as ML is the science of developing algorithms and statistical models for computer systems that can be used to create complex applications that can accurately classify and predict various types of data [3]. ML is a branch of artificial intelligence (AI) that is useful for creating models [5]. ML is also a technique or method used to learn extensive data and help make predictions [6]. ML learns from data distribution to make judgments about incoming data. The model or classifier can determine sales predictions based on the input features. The various processing methods provided by Machine Learning ML can be used effectively and affordably to forecast future sales. These ML algorithms are categorized into Reinforcement, Unsupervised, Semi-Supervised, and Supervised Learning [7]. ML works in three main parts: the decision-making process, error function, and model optimization process.

To increase the accuracy of ML results, you can use hyperparameter tuning to get the best parameters so that model accuracy increases. Hyperparameter tuning is finding the best set of hyperparameters from an ML model to get the maximum evaluation score [8]. One hyperparameter tuning that is relatively new and has advantages compared to other hyperparameter tuning is optuna hyperparameter tuning. Optuna is a hyperparameter tuning software framework specifically designed for machine learning, which is helpful in efficiently optimizing and finding more effective hyperparameters. It also has a pruning strategy to save computing resources and time [9].

Several researchers have researched sales predictions using ML. G. Beher and N. Nain 2020 researched sales forecasting or predictions using

ML algorithms, namely Linear Regression, Decision Tree, Ridge Regression, XGBoost with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) matrices to measure model performance and the value obtained is that the XGBoost model outperforms other methods where XGBoost obtains the lowest error value, namely the RMSE value is 1052 [10]. This research will become the primary reference for the research carried out. It is hoped that the results of this research will provide benefits to help accurately predict product level sales or product value in a store, thereby producing better accuracy values and reducing the risk of financial loss.

2. RESEARCH METHODOLOGY

2.1. Research Flow Scheme

Several stages were carried out in this research. These steps can be depicted in Figure 1 as follows:

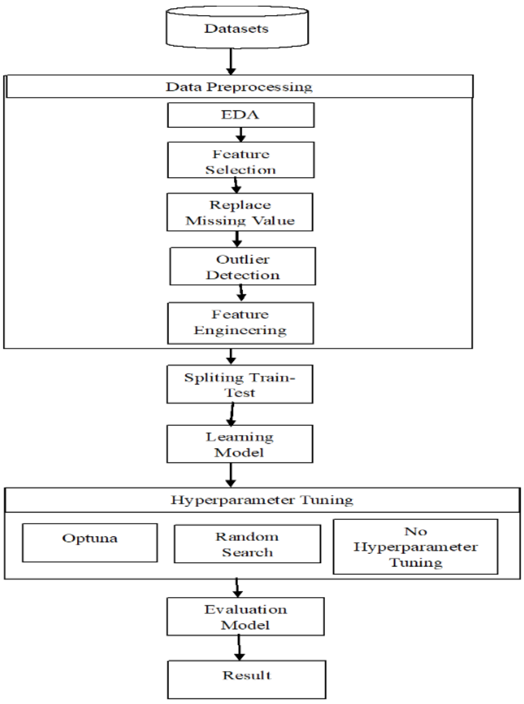

Figure 1. Research Methodology

2.2. Data Collection

In this research, data collection was carried out by searching for data secondaryly, namely publicly. The public data taken is data regarding product sales at Big Mart stores. This public data was taken from the online site at https://www.kaggle.com/datasets/brijbhushannanda1979/bigmart-sales-data/data. Product sales data at Big Mart stores consists of 12 attributes, 8523 train data, and 5681 test data.

2.3. Data Preprocessing

There are several stages in the data preprocessing process carried out in this research. Namely, the first stage is the Exploratory Data Analysis (EDA) process. EDA is a process carried out in data preprocessing that is applicable to help produce analysis results that help find patterns in data [11]. Then, the feature selection process, replacement of missing values, and feature engineering are carried out. The stages in the preprocessing process are carried out to provide convenience in the training data experimentation process to produce better model predictions. They can provide prediction results with smaller model error values.

2.4. Feature Selection

The process after EDA data preprocessing is to carry out feature selection. In this process, choosing what features will be used in the prediction process is helpful because not all features are relevant to the problem at hand. If these features are still used, they will affect the prediction results, so it is necessary to select existing features and lacking features. Influential should be removed [12]. The feature selection used in this research is using a correlation matrix. Matrix correlation functions to investigate the relationships that exist between numerical variables in a dataset [13].

2.5. Replace Missing Value

The replace missing value stage is the stage where the data contains NaN or empty values. Therefore, it is necessary to fill in NaN values in the empty attributes [14]. Attributes with missing values or NaN are not removed because they are a better choice. After all, data is well-spent.

2.6. Outlier Detection

Outlier detection will try to catch cases showing significant deviations from the existing majority pattern [15]. The outlier detection process is a process that is not easy to carry out because not all existing values are identified as outliers, so a statistical method is needed to identify and eliminate data that is in the abnormal range in the data set.

2.7. Feature Engineering

Feature engineering is necessary because the features in the data greatly influence the model performance results [16]. In the engineering process, an engineering process is carried out on several features in the dataset, namely the item_fat_content attribute because this attribute has upper and lower case letters, so it needs to be combined and added categories to item_fat_content, namely non-edible. Label encoding is carried out in the outle_size, item_fat_content, outlet_type, and outlet_location_type attributes, namely, changing non-numerical data into numeric value data. In the item_identifier attribute, the attribute is changed to item_identifier_categories, and products are categorized into three categories: food, drinks, and non-consumable goods [10]. Then, one hot encoding process will be carried out on the item_type, item_identifier_categories, and outlet_identifier tables.

2.8. Splitting Train-Test

At the splitting train and test stage, the data is divided into two parts based on the split selected as training data and the rest as testing data. Train data functions to fit the model in the ML process, while test data functions to evaluate the results of the fit model in the ML [17]. The split used in this research is 80:20.

2.9. Learning Model

After going through the data preprocessing process and splitting, the dataset is tested using an ML algorithm at the learning model stage.

2.10.Evaluation Model

After the testing process uses the ML algorithm, it will produce a prediction value, namely displaying the error value. The method for displaying the error value uses evaluation methods, namely RMSE, MAE, R2 Score, CV Score Mean, and Std.

2.11.Results

After experiments and trials, the best model will be produced, with the smallest error value obtained compared to reference research.

3. RESULTS AND DISCUSSION

In this research, several experiments were carried out, namely as follows:

3.1 ML Algorithm

In the first experiment, ML algorithm experiments were carried out without hyperparameter tuning, and no outlier detection process was carried out during data preprocessing. The following are the results of using the ML algorithm without hyperparameter tuning and without outlier detection in the data preprocessing process, which can be seen in Table 1.

TABLE I. ML ALGORITHM TEST RESULTS

Model	**RMSE**	**MAE**	**R2**	**CV Score Mean**	**CV Score Std**
XGB Regressor	1185	839	0.519	1186	46.397
Random Forest Regressor	1152	817	0.546	1158	39.674
LGBM Regressor	1113	781	0.576	1114	38.192
Lasso	1130	851	0.563	1129	42.014
AdaBoost Regressor	1192	927	0.514	1189	83.345
Decision Tree Regressor	1571	1097	0.155	1559	50.06
CatBoost Regressor	1115	789	0.574	1125	42.539
Ridge	1131	852	0.562	1130	41.488
KNeighbors Regressor	1242	898	0.472	1236	40.869
SVR	1551	1158	0.177	1550	61.422
ElasticNet	1245	935	0.47	1238	50.334
Extra Trees Regressor	1224	860	0.487	1253	43.014
MLP Regressor	1155	875	0.544	1156	44.41
Huber Regressor	1150	854	0.547	1140	46.655
RANSACRegressor	1287	941	0.433	1280	63.43

The first experiment using a regression algorithm from 15 ML algorithms without using hyperparameter tuning and without carrying out an outlier detection stage during preprocessing, produced relatively high error values, namely an average RMSE value above 1100, an average MAE value above 780, an average R2 value -average produces a value close to 1. Each algorithm's mean CV Score value is still above 1000, while the CV Score Std value has a reasonably high error value above 35.

3.2 ML Algorithm+Hyperparameter Tuning Random Search

TABLE II. ML ALGORITHM+*HYPERPARAMETER TUNING RANDOM SEARCH TEST RESULTS*

Model	RMSE	MAE	R2	CV Score Mean	CV Score Std
XGB Regressor	1093	776	0.591	1090	39.099
Random Forest Regressor	1090	772	0.593	1080	41.513
LGBM Regressor	1087	776	0.595	1077	40.920
Lasso	1130	851	0.563	1129	42.014
AdaBoost Regressor	1125	835	0.567	1119	40.454
Decision Tree Regressor	1109	794	0.579	1104	41.972
CatBoost Regressor	1276	996	0.443	1278	48.696
Ridge	1131	852	0.562	1130	41.488
KNeighbors Regressor	1128	823	0.564	1136	38.979
SVR	2687	2395	-1.468	2644	50.568
ElasticNet	1130	850	0.563	1129	42.095
Extra Trees Regressor	1081	764	0.600	1080	40.250
MLP Regressor	1102	781	0.584	1094	38.241
Huber Regressor	1147	847	0.550	1142	46.597
RANSACRegressor	1232	936	0.481	1270	49.118

In the second experiment, experiments were carried out using the ML algorithm, namely the regression algorithm, and hyperparameter tuning using random search without outlier detection. The results of the second experiment produced the lowest error value obtained by the extra trees regressor algorithm. However, the error value was still very high. Namely, the average error value of the 15 algorithms used, namely the RMSE error value, was still above 1000; therefore, it was necessary to try again.

3.3 ML Algorithm+*Hyperparameter Tuning Optuna*

TABLE III. ML ALGORITHM+*HYPERPARAMETER TUNING OPTUNA TEST RESULTS*

Model	RMSE	MAE	R2	CV Score Mean	CV Score Std
XGB Regressor	1096	774	0.589	1086	41.836
Random Forest Regressor	1095	774	0.59	1087	40.149
LGBM Regressor	1094	775	0.59	1087	39.864
Lasso	1130	851	0.563	1129	42.014
AdaBoost Regressor	1131	827	0.562	1118	44.5
Decision Tree Regressor	1710	1365	-4.421	1619	86..594
CatBoost Regressor	1187	913	0.518	1186	50.272
Ridge	1131	852	0.562	1130	41.488
KNeighbors Regressor	1247	940	0.468	1251	55.098
SVR	3406	3101	-2.965	3356	80.071
ElasticNet	1288	965	0.432	1279	51.403
Extra Trees Regressor	1087	770	0.596	3356	80.071
MLP Regressor	1141	848	0.555	1119	37.485
Huber Regressor	1233	921	0.48	1220	48.462
RANSACRegressor	1339	992	0.387	1267	56.992

In the third experiment, experiments were carried out using the ML algorithm, namely the regression algorithm used, and hyperparameter tuning was carried out using Optuna without outlier detection. The results of the third experiment produced relatively high error values. The results of the third experiment showed that the best algorithm was a random forest regressor. However, the error value was still high, namely the RMSE value above 1000; for this reason, it was necessary to test again.

3.4 ML Algorithm+ Outlier Detection

TABLE IV. ML ALGORITHM+*OUTLIER DETECTION TEST RESULTS*

Model	RMSE	MAE	R2	CV Score Mean	CV Score Std
XGB Regressor	1092	779	0.46	1051	36.809
Random Forest Regressor	1047	752	0.504	1024	29.536
LGBM Regressor	1013	730	0.536	989	27.474
Lasso	1014	779	0.534	1002	34.656
AdaBoost Regressor	1014	760	0.535	1002	31.767
Decision Tree Regressor	1428	1016	0.077	1405	43.419
CatBoost Regressor	1023	736	0.527	992	30.867
Ridge	1016	781	0.533	1003	34.431
KNeighbors Regressor	1119	821	0.434	1102	29.85
SVR	1332	1021	0.197	1366	56.469
ElasticNet	1098	837	0.454	1116	40.496
Extra Trees Regressor	1111	802	0.441	1104	39.756
MLP Regressor	1015	777	0.533	1057	31.827
Huber Regressor	1017	775	0.532	1007	35.148
RANSACRegressor	1081	826	0.471	1100	33.074

In the fourth experiment, experiments were carried out using the ML algorithm, namely the regression algorithm, and in the preprocessing process, the outlier detection stage was carried out. The results of the fourth experiment produced better values. Namely, the error value decreased compared to the previous experiment. So, by detecting outliers in the dataset, you can improve the results of the ML algorithm.

3.5 ML Algorithm+ Hyperparameter Tuning Random Search+Outlier Detection

TABLE V. ML ALGORITHM+ *HYPERPARAMETER TUNING RANDOM SEARCH*+DETEKSI *OUTLIER TEST RESULTS*

Model	RMSE	MAE	R2	CV Score Mean	CV Score Std
XGB Regressor	983	717	0.563	971	31.636
Random Forest Regressor	976	713	0.569	965	30.97
LGBM Regressor	972	708	0.572	961	31.501
Lasso	1014	779	0.534	1002	34.656
AdaBoost Regressor	1006	757	0.542	1001	31.646
Decision Tree Regressor	983	724	0.562	990	31.025
CatBoost Regressor	1122	886	0.431	1132	39.157
Ridge	1016	781	0.533	1003	34.431
KNeighbors Regressor	1030	760	0.52	1014	28.347
SVR	1849	1605	-0.544	1872	37.769
ElasticNet	1014	779	0.534	1002	34.716
Extra Trees Regressor	970	707	0.574	961	31.823
MLP Regressor	988	733	0.558	978	42.778
Huber Regressor	1021	775	0.529	1010	36.417
RANSACRegressor	1129	876	0.423	1117	42.741

The fifth experiment produced quite good values when the fifth experiment used hyperparameter tuning random search and added an outlier detection process during the preprocessing process. However, when using random search hyperparameter tuning, the SVR algorithm experienced a decrease.

3.6 ML Algorithm ML+Hyperparameter Tuning Optuna+Outlier Detection

TABLE VI. ML ALGORITHM+*HYPERPARAMETER TUNING OPTUNA*+OUTLIER DETECTION TEST RESULTS

Model	RMSE	MAE	R2	CV Score Mean	CV Score Std
XGB Regressor	968	708	0.576	963	31.991
Random Forest Regressor	977	712	0.568	967	29.373
LGBM Regressor	971	709	0.573	962	32.617
Lasso	1014	779	0.534	1002	34.656
AdaBoost Regressor	1006	756	0.542	1002	32.164
Decision Tree Regressor	1488	1211	-0.0003	1500	53.811
CatBoost Regressor	1051	818	0.5	1058	42.403
Ridge	1016	781	0.533	10033	34.431
KNeighbors Regressor	1094	830	0.458	1122	35.211
SVR	1895	1650	0.622	1929	37.545
ElasticNet	1132	862	0.42	1154	41.453
Extra Trees Regressor	973	707	0.572	1929	37.545
MLP Regressor	998	745	0.549	1006	39.41
Huber Regressor	1058	807	0.494	1072	39.897
RANSACRegressor	1094	851	0.458	1122	40.743

In the sixth experiment, the ML algorithm uses optuna hyperparameter tuning and an outlier detection process is carried out during data preprocessing. There is a pretty good decrease in the error value compared to the fifth experiment, where the sixth experiment produces the algorithm with the lowest error value, namely the XGBoost algorithm. The error value in the XGBoost algorithm decreased by 15 in the RMSE value, and the MAE error value decreased by 9, while the R2 value also improved, namely 0.576, and the CV Score mean and Std values were also better than in the fifth experiment.

After generating predicted values using the R2, MAE, RMSE, CV Score Mean, and Std evaluation matrices from each ML algorithm with six trials, it was found that XGBoost Regressor obtained the best algorithm with optuna hyperparameter tuning and when the outlier detection process was carried out during the data preprocessing process.

The research results produced an RMSE value of 968 and an MAE value of 713, and this shows that the research results showed that the RMSE error value decreased by 84 and the MAE value decreased by 26.03 from the reference paper.

TABLE VII. PREDICTION OF RESULTS USING THE REFERENCE MODEL

Nama	*RMSE*	*MAE*
ML Algorithm +*Optuna*+Outlier Detection	968	713
Reference Paper	1052	739.03

From several experiments that have been carried out on several ML algorithms, it was found that the best results used to predict the level/value of product sales in stores were experiments with the XGBoost Regressor algorithm using Optuna hyperparameter tuning with outlier detection producing quite good error values compared to the reference paper. That is, it produces an RMSE error value of 968, reducing the error value by 84.

Figure 2 below shows several examples of predictions of sales levels or values using Big Mart sales data.

	Outlet_Type	Outlet_Size	Item_Weight	Item_Fat_Content	Outlet_Location_Type	Outlet_Age	Item_Outlet_Sales
0	1	2	20.750000	0	0	25	1641.148926
1	1	2	8.300000	2	1	17	1428.293579
2	0	2	14.600000	1	2	26	821.942139
3	1	2	7.315000	0	1	17	2456.098633
4	3	2	12.857645	2	2	39	4405.342773

Figure 2. Prediction Of Product Sales Levels

The results of the prediction of the level or value of product sales, where from the 5 examples of predicted data, the outlet_type, which is a grocery store type in the table, has a value of 0, has the lowest sales value compared to supermarket type, stores, even though the store is old. Meanwhile, sales in supermarkets are higher, and even though the size of the shop is medium, sales are still significant. This is due to one of the shop locations, where residents in tier 3 prefer supermarkets to grocery shops.

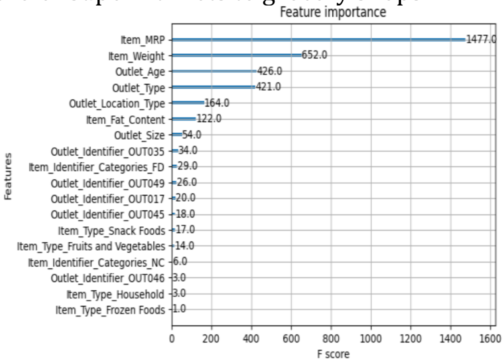

Figure 3. Feature Importance Value

Figure 3 shows the value of feature importance using the XGBoost algorithm model with optuna tuning hyperparameters because it has the lowest prediction error value. The results of the feature importance value show that the item_MRP attribute is the highest. This shows that the item_MRP attribute has the most influence on predicting a store's value or level of product sales. Then, continue with item_weight, the second most influential attribute in predicting sales level or value. Other influential attributes are outlet_age, outlet_type, outlet_location_type, item_fat_content, outlet_size, outlet_identifier, item_identifier, and item_type.

4. Conclusions and Suggestions

The choice of ML algorithm determines the results of good predictions from existing data and is also determined by the parameters used. Optuna hyperparameter tuning can reduce the error value in the XGboost regressor algorithm and is the best algorithm result compared to 15 other experimental algorithms. Outlier detection in the preprocessing process can improve the results of the ML algorithm. Evaluation results from research that has been carried out using the metrics RMSE, MAE, R2, CV Score Mean, and Std show that by using an ML algorithm model with optimal hyperparameter tuning and carrying out an appropriate preprocessing process, namely by carrying out outlier detection, it is possible to find a model with the correct error value—decreased compared to algorithm experiments without outlier detection. The influential features from the research results that have been carried out are the item_MRP, item_weight, outlet_age, and outlet_type features. So, the value or level of product sales is influenced by the product's retail price.

The suggestions from this research are to re-adjust each parameter of the ML algorithm used to improve the performance of the best algorithm in predicting the level or value of product sales in stores and develop the use of other data preprocessing processes, such as for handling outlier data, and also carry out the development of feature selection in selecting attributes that have a strong influence on the algorithm model used and do not reduce the results of the algorithm used, as well as carry out further development of the use of hyperparameter tuning to improve the results of the algorithm used.

5. ACKNOWLEDGMENTS

The author takes this opportunity to thank the Chancellor of Nusa Mandiri University, Mrs Prof. Dr Dwiza Riana, S.Si, MM, M.Kom, Deputy Chancellor of Nusa Mandiri University, Head of the Computer Science Study Program at Nusa Mandiri University, Mr Dr Muhamad Haris, M.Eng as the thesis supervisor, his beloved parents and older siblings always provide both material and moral support to the author and all parties who cannot be mentioned one by one.

Reference:

[1] X. Dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," *2021 IEEE Int. Conf. Consum. Electron. Comput. Eng. ICCECE 2021*, no. Iccece, pp. 480–483, 2021, doi: 10.1109/ICCECE51280.2021.9342304.

[2] D. Preethi and V. S, "An Analysis of Machine Learning Algorithms to Predict Sales," *Int. J. Sci. Res.*, vol. 11, no. 6, pp. 462–466, 2022, doi: 10.21275/SR22601144946.

[3] J. Wang, "A hybrid machine learning model for sales prediction," *Proc. - 2020 Int. Conf. Intell. Comput. Human-Computer Interact. ICHCI 2020*, pp. 363–366, 2020, doi: 10.1109/ICHCI51889.2020.00083.

[4] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," *Proc. 2018 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2018*, pp. 160–166, 2018, doi: 10.1109/CSITSS.2018.8768765.

[5] N. Giarsyani, "Komparasi Algoritma Machine Learning dan Deep Learning untuk Named Entity Recognition : Studi Kasus Data Kebencanaan," *Indones. J. Appl. Informatics*, vol. 4, no. 2, p. 138,

2020, doi: 10.20961/ijai.v4i2.41317.

[6] Muflih Ihza Rifatama, Mohammad Reza Faisal, Rudy Herteno, Irwan Budiman, and Muhammad Itqan Mazdadi, "Optimasi Algoritma K-Nearest Neighbor Dengan Seleksi Fitur Menggunakan Xgboost," *J. Inform. dan Rekayasa Elektron.*, vol. 6, no. 1, pp. 64–72, 2023, doi: 10.36595/jire.v6i1.723.

[7] R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. ICICCS 2018*, no. Iciccs, pp. 945–949, 2019, doi: 10.1109/ICCONS.2018.8663155.

[8] L. Owen, *Hyperparameter Tuning with Python*. Birmingham: Packt Publishing Ltd, 2022.

[9] A. Van Wyk, *Machine Learning with LightGBM and Python*. Birmingham: Packt Publishing Ltd, 2023.

[10] G. Behera and N. Nain, "A comparative study of big mart sales prediction," *Commun. Comput. Inf. Sci.*, vol. 1147 CCIS, no. March, pp. 421–432, 2020, doi: 10.1007/978-981-15-4015-8_37.

[11] E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," *Pros. Nas. Rekayasa Teknol. Ind. dan Inf. XIV Tahun 2019*, vol. 2019, no. November, pp. 263–269, 2019, [Online]. Available: http://journal.itny.ac.id/index.php/ReTII

[12] D. Kurniawan, *Pengenalan Machine Learning dengan Python*. Jakarta: PT Elex Media Komputindo, 2020.

[13] A. F. A. Naibaho and A. Zahra, "Prediksi Kelulusan Siswa Sekolah Menengah Pertama Menggunakan Machine Learning," *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3, 2023, doi: 10.23960/jitet.v11i3.3056.

[14] B. Purnama, *Pengantar Machine Learning Konsep dan Praktikum Dengan Contoh Latihan Berbasis R dan Python*. Bandung: Informatika Bandung, 2019.

[15] S. Gracia, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Granada: Springer, 2014.

[16] I. Daqiqil Id, *Machine Learning Teori, Studi Kasus dan Implementasi Menggunakan Python*. Pekanbaru: UR PRESS, 2021.

[17] D. A. Rhamadhani, E. Eklezia, and D. Saputri, "Analisa Model Machine Learning dalam Memprediksi Laju Produksi Sumur Migas 15/9-F-14H," *J. Sustain. Energy Dev.*, vol. 1, no. 1, pp. 48–55, 2023.