Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 6, No.2, September 2023, pp. 252 – 261
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Prediction Model of Revenue Restaurant's Business Using Random Forest

[1]Erfan Ainul Yakin, [2]Ririen Kusumawati, [3]Usman Pagalay
[1,2,3]Faculty of Science and Technology, Program Specification for Master Study in Computer Science (Magister Informatika) Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia
Email: [1]erfan88y@gmail.com, [2]ririen.kusumawati@ti.uin-malang.ac.id, [3]usman@mat.uin-malang.ac.id.

Article Info
Article history:
Received Jul 4th, 2023
Revised Aug 17th, 2023
Accepted Sep 6th, 2023

Keyword:
Machine Learning
Prediction
Random Forest
Restaurant's Business
Revenue

ABSTRACT
This research was conducted to predict the level of revenue from the Soto Kwali Pak Wasis restaurant business using Machine Learning. The Random Forest method was chosen because it can predict optimal and fast results with low hardware requirements. Prediction Model results using the Random Forest method resulted in an average accuracy value of 75.4% from a combination of 4 experiments. Thus, the Random Forest method is one of the flexible algorithms and is very suitable for predicting revenue in the Soto Kwali Pak Wasis restaurant business because of its good speed, high accuracy, and requires lower costs.

Corresponding Author:
Erfan Ainul Yakin,
Faculty of Science and Technology,
Program Specification for Master Study in Computer Science (Magister Informatika),
Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia,
Jl. Gajayana No.50, Dinoyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65144.
Email: erfan88y@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v6i2.24984

1. INTRODUCTION

Small-scale business activities in the field of management need to be evaluated, especially on quality. Quality can guarantee the improvement of business activities, especially in the evaluation of the administrative system of a business that has been operating for a long time. The difficulty of developing small business activities is caused by entrepreneurs focusing too much on external things or customer requests so the evaluation of management and administration is often neglected [8].

Soto kwali pak wasis restaurant is a small-scale business that began to grow rapidly in Surabaya and Sidoarjo locations. Business management and administration used in this restaurant still uses manual methods and are done by the restaurant owner himself. In the manual record, transactions that occur in this restaurant, which consists of several branches, are so high that restaurant owners work with application developers to create a simple administration system. The administrative system that has been applied to this restaurant is limited to only presenting information on the number of foods sold each day and the nominal money received from these sales, so evaluating the development of this restaurant is still difficult to do [16].

Evaluation of raw data taken from the Soto Kwali Pak Wasis restaurant administration system using Machine Learning technology. Machine learning performs calculations from data on the amount of food sold at a certain time and the results of these calculations will be used to generate a Prediction Model [7]. The Prediction Model uses level categories to facilitate the evaluation of income received within a certain time so that it can be used as a reference for determining business development in the future [4].

Today's enterprise field is in dire need of prediction technology. Prediction by measuring the value of a business's revenue can function in business development and remain afloat and prosperous [15]. Sales predictions using Machine Learning are essential for better growth and sales. Machine Learning and Data Science technologies make it easy to determine the algorithm or method used to determine model predictions [2].

The most popular Machine Learning algorithm that is often used for prediction is Random Forest. Random Forest uses Gini index calculations and variable evaluation fund entropy to produce accuracy [17]. The stages of the Random Forest algorithm are so many that they require a high level of calculation which results in requiring good hardware resources. The heavy performance load of the Random Forest algorithm calculation results in a very good level of accuracy so it is suitable for application in model predictions [9].

Research conducted by Ghorbani *et al*. (2020) on the topic of predicting student performance using the Random Forest method resulted in an accuracy rate of 74%. The advantages of the Random Forest algorithm in addition to high accuracy are also a fast calculation process and do not require high hardware [11].

Food price prediction conducted by Sahinbas *et al.* (2022) with the topic of food price prediction in Istanbul restaurants with the Random Forest method achieved an accuracy rate of 98%. The Random forest method in this study produces a very low *MAE* so it is very good for making predictions in the restaurant field [6].

In this study, the prediction model for the soto kwali pak wasis restaurant business was optimized with the Random Forest algorithm to produce good prediction results. Prediction results that are close to accurate can be used for business development and evaluation so that steps for business preparation become more mature [3].

2. RESEARCH METHOD

Procedural research in this study is carried out in several stages, each stage is still related to each other and although the initial stage has not been reached optimally, the next stage can still be done. The flow of procedure research is shown in Figure 1.

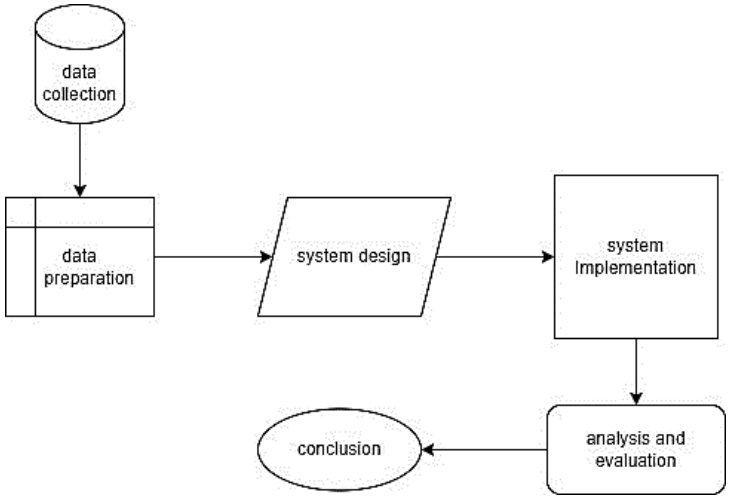

Figure 1. Research Design

The data collection uses data on sales of food items per day in Pak Wasis' soto kwali restaurant business which is primary data and taken in April 2023. The next stage is to do data preparation from the data that has been collected into data that is ready for processing, namely by selecting several attributes independent and dependent on the data. The independent attribute is the names of the foods or beverages sold and the dependent attribute is the level category of the income level [10].

Design system with Random Forest algorithm based on CART Classification and Regression Trees method using gini index weights as calculation parameters [20]. In the system design, data will be separated into training data and testing data. Training data will be studied and evaluated by algorithms that are used as a reference in carrying out the data testing process. The Random Forest process will count and create a large number of trees and from many trees will be voted on the value produced by the tree [19]. Gini calculation according to the equation (1):

$$Gini(S) = \sum_{j=1}^{n} P_j^2$$
$$Gini_A(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2) \quad (1)$$

Where P_j is the data chance of the variable *j*.

After the Random Forest calculation results are obtained, the next process is evaluated using *MAE* & *MAPE* and matches with real conditions using confusion on matrix [12]. The output of the system is a prediction level consisting of high, standard and low. The design system is shown in figure 2

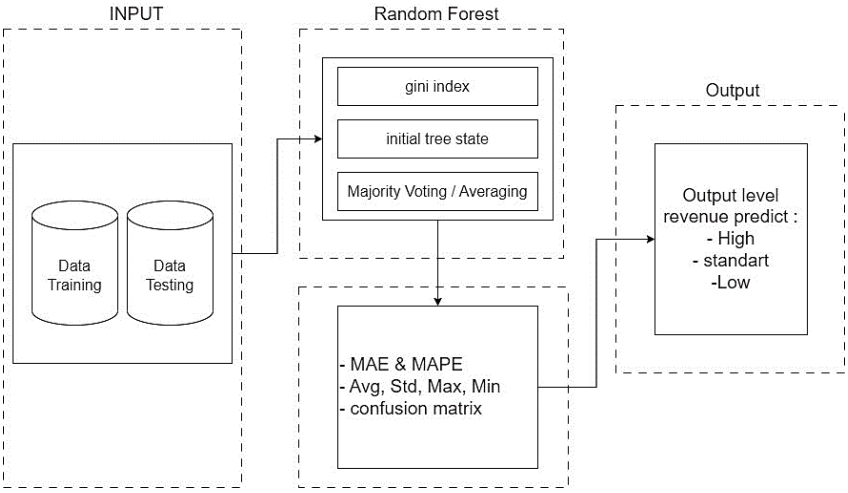

Figure 2. Design System

The implementation system uses Python programming language by doing 4 data test scenarios. From each scenario will be applied tree state 50, 100, 500, and 1000. The results will be calculated with *MAE and* MAPE and evaluated using a confusion matrix for accuracy to display the prediction results properly [14]. Data trial scenarios are shown in Table 1.

Table 1. Trial Scenario

Train Name	Compare Train : Test	Tree State
P1	90:10	50,100,500,1000
P2	80:20	50,100,500,1000
P3	70:30	50,100,500,1000
P4	60:40	50,100,500,1000

The process flow of the Random Forest algorithm starts with data input and continues with separate training data and test data. The next process is to process data using the tree method which is repeated to n [13]. If the accuracy value is above the lowest tolerance, then the next step is evaluation using the *MAE* and *MAPE* methods. The flowchart diagram is shown in Figure 3.

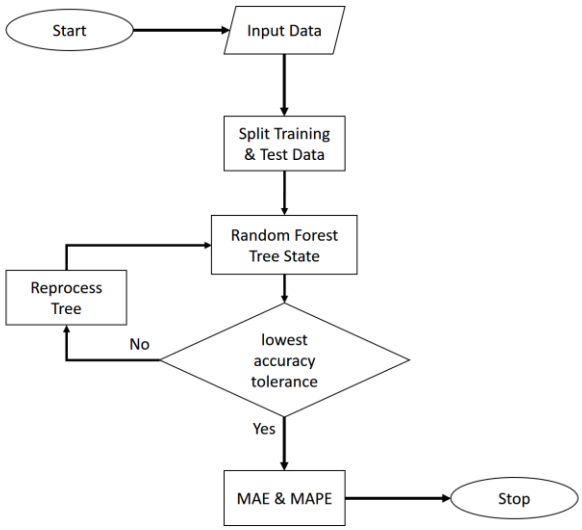

Figure 3. Flowchart Random Forest

3. RESULTS AND ANALYSIS

The implementation of the system is applied to computer devices with core i5 specifications, 8GB RAM, and SSD using Python programming language. The 182 data were formed using CSV format and explored with detailed variables as in Table 2.

Table 2. Variabel dataset

No	Feature	Type	Variabel
1	Date	Datetime	Independent
2	Drinking Water(WTR)	Numeric	Independent
3	Lemon Warm/Cold(JRK)	Numeric	Independent
4	Chicken Soup Big(SAB)	Numeric	Independent
5	Chicken Soup Small(SAK)	Numeric	Independent
6	Meat Soup Big(SDB)	Numeric	Independent
7	Meat Soup Small(SDK)	Numeric	Independent
8	Meatball Tofu(TBS)	Numeric	Independent
9	Tea Warm/Cold(TEA)	Numeric	Independent
10	Mendoan Tempeh(TMD)	Numeric	Independent
11	Level	Text	Dependet

In the process of applying the Random Forest algorithm, the dataset that has been trained can be known as the features that have the most influence on the prediction process. The feature with the name Tea Warm/Cold (TEA) is the feature that gets the highest score as shown in Figure 4.

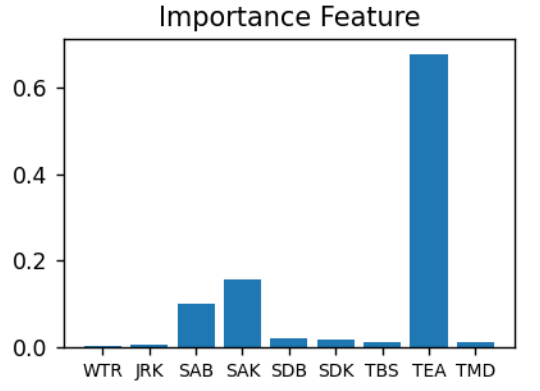

Figure 4. Importance Feature

Implementation will be carried out in 4 scenarios as described in the previous chapter. This implementation aims to determine the level of accuracy and time efficiency in carrying out the calculation process so that it is expected that the application of the method to the ongoing business can be done with little cost but good prediction results.

3.1. Scenario 1

In scenario 1, the dataset uses a ratio of 90:10 with a state random forest configuration using levels 50, 100, 500, and 1000. Scenario 1 conducts several experiments aimed at obtaining MAE, MAPE, accuracy, and duration values. The results of scenario 1 are shown in Table 3.

Table 3. Result Scenario 1

Tree State	MAE	MAPE	Accuracy	Duration
50	0,00526	1,85	0,84	12s
100	0,00421	3,16	0,79	15s
500	0,00526	4,32	0,65	28s
1000	0,00598	23,70	0,48	39s

The results of scenario 1 show that if the comparison of test data is 10% and training data is 90%, the most optimal result is at tree state 50. The best accuracy results are 84% with an execution time of 12 seconds. The error rate of the calculation is still very small at below 0.1 and the difference between real conditions and predictions is 1.85%. A comparison of real conditions and predictions is depicted in Figure 5.

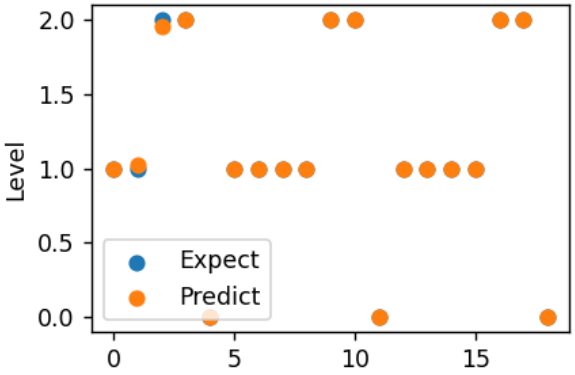

Figure 5. Plot expect and predict value scenario 1 best value

In Figure 5 real conditions are depicted with blue plots and predicted conditions are depicted with yellow plots. Real conditions and predictions are almost accurate in the range of day 4 to day 18. While the slightly distorted prediction conditions are shown on day 3 which depicts blue and yellow plots containing and on day 2 also blue and yellow plots intersecting. Comparison of real conditions and more detailed predictions in Figure 6.

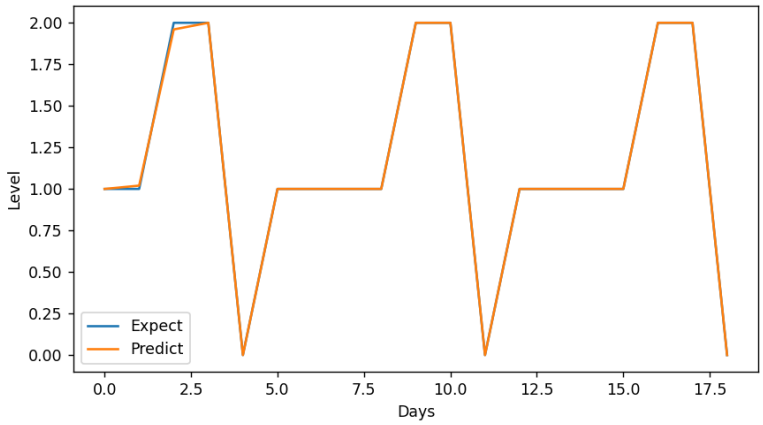

Figure 6. Expect and Predict average scenario 1 best value

Figure 6 on days 4 to 18 shows the results of the yellow prediction line that almost does not deviate from blue, the yellow line that deviates on days 2 and 3. So it can be concluded from scenario 1 with tree state 50 shows excellent prediction results with low error values and faster processing times.

3.2. Scenario 2

Scenario 2, applied trial with a ratio of 80:20. Scenario 2 also applies the experiment several times and produces MAE, MAPE, accuracy, and duration values. The results of scenario 2 are shown in Table 4.

Table 4. Result in Scenario 2

Tree State	MAE	MAPE	Accuracy	Duration
50	0,02000	1,55	0,72	15s
100	0,02028	1,78	0,72	22s
500	0,02311	10,28	0,59	39s
1000	0,02094	13,65	0,53	66s

Scenario 2 shows the most optimal result remains at tree state 50 but also at state 100. The best accuracy results are 72% with an execution time of 15 seconds but in a longer duration state of 22 seconds. The error rate of the calculation is still very small at below 0.1 and the difference between real conditions and predictions is still below 2%. A comparison of real conditions and predictions is depicted in Figure 7.

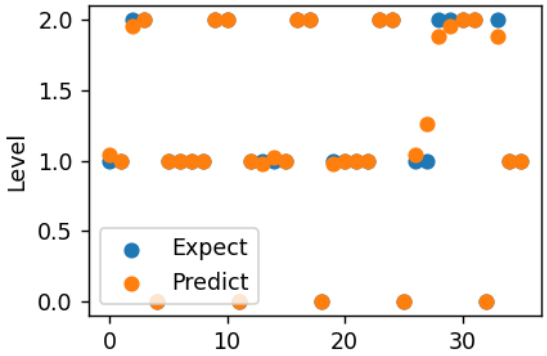

Figure 7. Plot expect and predict value scenario 2 best value

In Figure 7 there are almost more real conditions and predictions than in scenario 1. Divergent prediction conditions were shown on days 1,2,26,28 yellow and blue plots intersecting each other and on days 27, and 32 yellow and blue plots diverged far thus showing lower prediction accuracy. Comparison of real conditions and more detailed predictions in Figure 8.

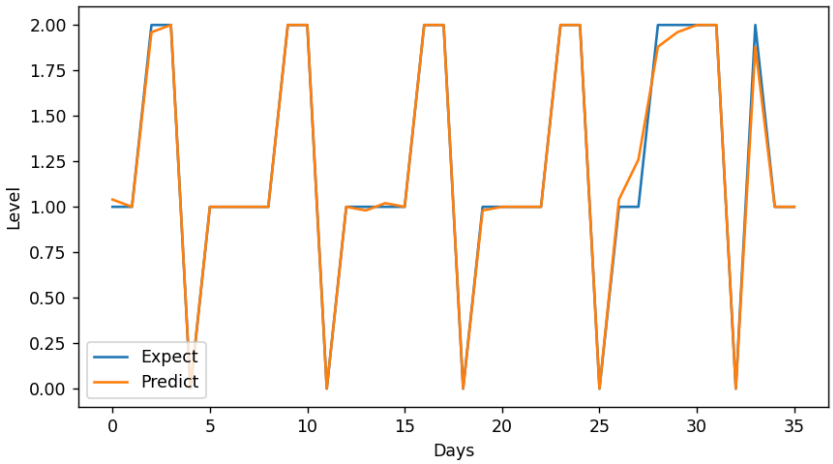

Figure 8. Expect and Predict average scenario 2 best value

Figure 8 on days 1, 2, 13, and 28 shows the results of real lines that deviate more with blue, and day 27 shows deviations from yellow and blue lines that are farther away, so it can be concluded that the error value in tree state 100 is higher. Tree state 100 vulnerable data still shows more yellow lines remain more dominant so the accuracy results are still above 70%. Scenario 2 with tree states 50 and 100 shows good prediction results and processing times are still relatively fast.

3.3. Scenario 3

Scenario 3, applying a test with conditions comparing train data and test data 70: 30. In scenario 3, several trials were carried out to get the optimal MAE, MAPE, accuracy, and d, duration values. The results of scenario 3 are shown in Table 5.

Table 5. Result Scenario 3

Tree State	MAE	MAPE	Accuracy	Duration
50	0,02227	16,10	0,75	15s
100	0,02236	42,31	0,69	20s
500	0,02260	50,76	0,56	54s
1000	0,02331	54,04	0,51	112s

Scenario 3 shows the most optimal results remain in tree state 50 with the condition of a greater number of test data. The best accuracy results are 75% with an execution time of 15 seconds. The error rate of the calculation is still very small, which is below 0.1. In scenario 3, the difference between real and predicted conditions starts to show a higher figure of 16%. A comparison of real conditions and predictions is depicted in Figure 9.

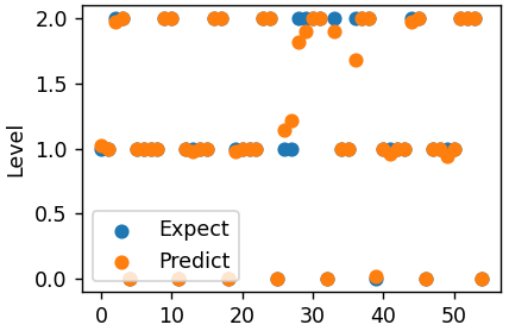

Figure 9. Plot expect and predict value scenario 3 best value

Scenario 3 shows a MAPE number of 16% because the real plot and prediction plot in Figure 9 are more different. There are approximately 10 different days between real plots and prediction plots. The yellow color plots show intersect with the blue color plots on days 1,3,12,20,37,39,42,45 and 49. While the yellow plots that deviate far are found on days 26, 27, 28, and 33 so the error rate is very high and the accuracy value tends to be low. Comparison of real conditions and more detailed predictions in Figure 10.

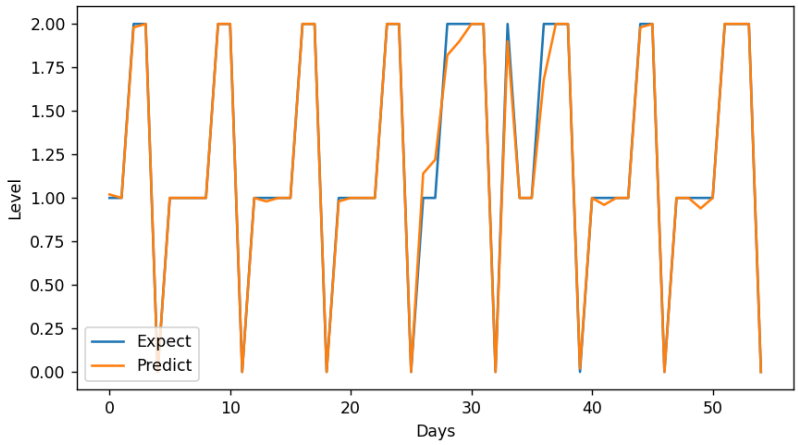

Figure 10. Expect and Predict average scenario 3 best value

Figure 10 shows the results of the prediction line are more dominant, but because the amount of test data used is 30%, the prediction results also experience a decrease in accuracy. Yellow lines that diverge far from blue lines are shown on days 27, 28, 29, and 38. While the yellow line that diverges not too far from the blue line is shown on days 1,3,12,20,37,42 and 45. Conditions in scenario 3 are still able to show good prediction results with a duration that is not too far from scenarios 1 and 2 but the error rate in scenario 3 is higher.

3.4. Scenario 4

Scenario 4, applies almost the same trial conditions of 60: 40. In scenario 4 for MAE, MAPE, accuracy, and optimal duration require calculation and longer duration. The results of scenario 4 are shown in Table 6.

Table 6. Result in Scenario 4

Tree State	MAE	MAPE	Accuracy	Duration
50	0,02043	42,54	0,74	17s
100	0,02056	43,78	0,72	22s
500	0,02106	54,36	0,57	65s
1000	0,02279	56,28	0,57	128s

Scenario 4 with tree state condition 50 shows the best results. The best accuracy results are 74% with an execution time of 17 seconds. The error rate of the calculation is still very small, which is below 0.1. In scenario 4, the difference between real and predicted conditions starts to show a very high figure of 42.54%. A comparison of real conditions and predictions is depicted in Figure 11.

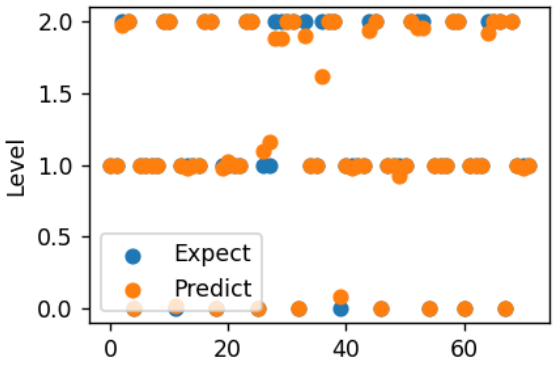

Figure 11. Plot expect and predict value scenario 4 best value

Scenario 4 comparison of real conditions and predicted conditions is very high, reaching MAPE 42.54% because the training data and test data are almost the same. There are approximately 30 different days between real plots and prediction plots. On days 3-5, 10-12, 20, 24-50, and 60-65 there are many wedges between yellow plots and blue plots. On days 25-28, the yellow plot deviates far from the blue plot so the error rate in scenario 4 is very high. Comparison of real conditions and more detailed predictions in Figure 12.

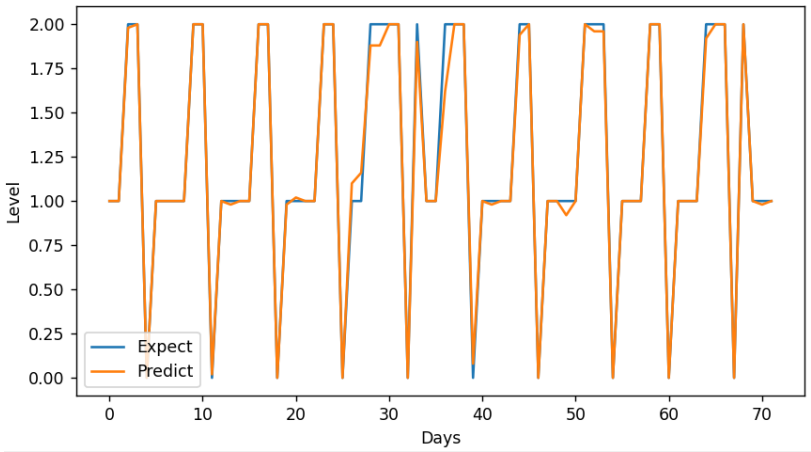

Figure 12. Expect and Predict average scenario 4 best value

Figure 12 real lines are more distractions so more blue lines appear in almost every vulnerable day. On days 26,27,35,50,53 and 65 showing the yellow line very far away from the blue line, this result causes a high error value. In test conditions, data with a ratio of 60:40 is not good for use as a prediction configuration.

All scenarios are summarized in Table 7, which shows the best results from experiments with different tree state configurations. The average of the 4 scenarios with 75.4% accuracy shows that Random forest is still very good to use for predictions.

Table 7. Summary Best Result All Scenario

Scenario	Tree State	MAE	MAPE	Accuracy	Duration
1	50	0,00526	1,85	0,84	12s
2	50	0,02000	1,55	0,72	15s
2	100	0,02028	1,78	0,72	22s
3	50	0,02227	16,10	0,75	15s
4	50	0,02043	42,54	0,74	17s

The best scenario compared to other scenarios is scenario 1 with a 90:10 comparison of train data and test data. For a small amount of training data less than 200 scenario 1 is optimally applied in the prediction model.

4. CONCLUSION

Prediction models are built using Random Forest and configured to be able to produce optimal values using tree state 50. Random Forest is an algorithm that can adapt to various types of data so that it is very good at accurately predicting data [1]. The application of prediction models with the random forest method in the

restaurant business can produce good analysis and requires simple hardware, so it is very suitable to be applied in the administration system [18]. Predictions that have been obtained can be used to prepare employee performance conditions and evaluate the business that has been running.

The test results of the Prediction model using Random Forest resulted in an average accuracy value of 75.4% with a tree state configuration of 50. The results of the comparison of predictions and real conditions are influenced by little or many test data, the more test data, the farther the prediction range with real conditions, and vice versa. The error rate resulting from the test is very low at below 0.1. Prediction of revenue levels using the Random Forest method at Soto Kwali Pak Wasis restaurant is very suitable to be applied with the aim of business evaluation in the future and can be prepared more carefully.

For future improvements, the researcher suggested that different things in the scope of data be taken in businesses with a larger scale and use more datasets and varied data attributes and need to experiment in another scope of business [5].

REFERENCES

[1] B. Farnham, S. Tokyo, B. Boston, F. Sebastopol, and T. Beijing, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION," Sebastopol, CA: O'Reilly Media, 2019, pp. 199–200.

[2] I. Kumar, J. Rawat, N. Mohd, and S. Husain, "Opportunities of Artificial Intelligence and Machine Learning in the Food Industry," J Food Qual, vol. 2021, 2021, doi: 10.1155/2021/4535567.

[3] K. Hakala et al., "Neural Network and Random Forest Models in Protein Function Prediction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 3, pp. 1772-1781, 1 May-June 2022, doi: 10.1109/TCBB.2020.3044230.

[4] K. Saetia and J. Yokrattanasak, "Stock Movement Prediction Using Machine Learning Based on Technical Indicators and Google Trend Searches in Thailand," International Journal of Financial Studies, vol. 11, no. 1, Mar. 2023, doi: 10.3390/ijfs11010005.

[5] K. Singh, P. M. Booma, and U. Eaganathan, "E-Commerce System for Sale Prediction Using Machine Learning Technique," in Journal of Physics: Conference Series, IOP Publishing Ltd, Dec. 2020. doi: 10.1088/1742-6596/1712/1/012042.

[6] K. ŞAHİNBAŞ, "Price Prediction Model for Restaurants In Istanbul By Using Machine Learning Algorithms," Ekonomi İşletme ve Maliye Araştırmaları Dergisi, vol. 4, no. 2, pp. 159–171, Aug. 2022, doi: 10.38009/ekimad.1148216.

[7] K. Zahoor, N. Z. Bawany, and S. Hamid, "Sentiment analysis and classification of restaurant reviews using machine learning," in Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ACIT50332.2020.9300098.

[8] M. Shoira, F. S. Usmanov, and B. S. Ubaydullayev, "Problems in the Implementation of Quality Management Systems in Small Business Enterprises," , Erb, vol. 7, pp. 54–57, Apr. 2022.

[9] Mr. T. Phase, "Predict the Level of Income using Random Forest Classifier," Int J Res Appl Sci Eng Technol, vol. 7, no. 12, pp. 558–561, Dec. 2019, doi: 10.22214/ijraset.2019.12090.

[10] P. Bajaj, R. Ray, S. Shedge, S. Vidhate, and D. Nikhilkumar, "SALES PREDICTION USING MACHINE LEARNING ALGORITHMS," International Research Journal of Engineering and Technology (IRJET), vol. 07, no. 06, 2020.

[11] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[12] S. D. A. Bujang et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," IEEE Access, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[13] S. Siddamsetty, R. Reddy Vangala, L. Reddy, and R. Vattipally, "Restaurant Revenue Prediction using Machine Learning," International Research Journal of Engineering and Technology, pp.2395-0056, 2021.

[14] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," J Big Data, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00516-9.

[15] U. M. Sirisha, M. C. Belavagi, and G. Attigeri, "Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison," IEEE Access, vol. 10, pp. 124715–124727, 2022, doi: 10.1109/ACCESS.2022.3224938.

[16] Umam, A. K., R. T. Ratnasari, and S. Herianingrum. "THE EFFECT OF MACROECONOMIC VARIABLES IN PREDICTING INDONESIAN SHARIA STOCK INDEX". Journal of Islamic Economics and Business, vol. 5, no. 2, Dec. 2019, pp. 223-40, doi:10.20473/jebis.v5i2.15031.

[17] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," Big Data Mining and Analytics, vol. 4, no. 2, pp. 116–123, Jun. 2021, doi: 10.26599/BDMA.2020.9020016.

[18] Y. Lin et al., "Revenue prediction for integrated renewable energy and energy storage system using machine learning techniques," J Energy Storage, vol. 50, Jun. 2022, doi: 10.1016/j.est.2022.104123.

[19] Y. R. Chen, J. S. Leu, S. A. Huang, J. T. Wang, and J. I. Takada, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," IEEE Access, vol. 9, pp. 73103–73109, 2021, doi: 10.1109/ACCESS.2021.3079701.

[20] Z. Jiang, "Prediction and Management of Regional Economic Scale Based on Machine Learning Model," Wirel Commun Mob Comput, vol. 2022, 2022, doi: 10.1155/2022/2083099.

BIBLIOGRAPHY OF AUTHORS

Erfan Ainul Yakin holds a Master of Informatics degree from UIN Malang University, Indonesia in 2022. Bachelor of Informatics degree from UMM Malang University, Indonesia in 2012. He can be contacted at email: erfan88y@gmail.com

Ririen Kusumawati Received the bachelor's degree from Universitas Brawijaya, in 1995, and the master's degree from Institut Teknologi Surabaya, in 2004, she received Doctoral degree in 2021 from Universitas Negeri Malang. She was a Lecture in informatics — engineering — with Universitas Islam Negeri Maualana Malik Ibrahim Malang. She can be contacted at email: ririen.kusumawati@ti.uin-malang.ac.id.

Usman Pagalay Received the bachelor's degree from Universitas Hasanuddin, in 1992, and the master's degree from Institut Teknologi Bandung, in 2003, he received Doctoral degree in 2012 from Universitas Brawijaya. He was a Lecture in informatics — engineering — with Universitas Islam Negeri Maualana Malik Ibrahim Malang. He can be contacted at email: usman@mat.uin-malang.ac.id.