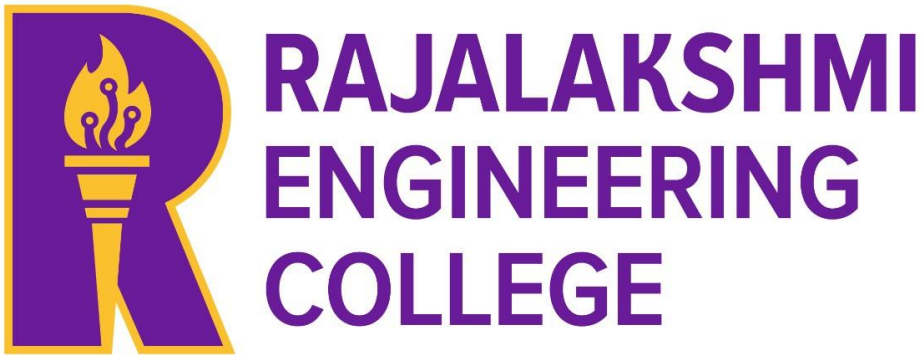


RAJALAKSHMI ENGINEERING COLLEGE

(An Autonomous Institution Affiliated to Anna University, Chennai)

PREDICITING FOOD PREPARATION TIME USING DATA SCIENCE

SUBMITTED BY
SARAVANA A (231801157)
PRIYAN N (231801130)

AD23532 – PRINCIPLES OF DATA SCIENCE
Department of Artificial Intelligence and Data Science
Rajalakshmi Engineering College, Thandalam
October 2025

BONAFIDE CERTIFICATE

NAME……

ACADEMIC YEAR……SEMESTER……. BRANCH ……..

UNIVERSITY REGISTER NO.

Certified that this is the Bonafide record of work done by the above student in the Mini Project titled " PREDICTING FOOD PREPARATION TIME USING DATA SCIENCE " in the subject AD23532 – PRINCIPLES OF DATA SCIENCE during the year 2025-2026

Signature of Faculty – in Charge

Submitted for the Practical Examination held on ----------

Internal Examiner External Examiner

S. No.	INDEX	Page No.

ABSTRACT

The accurate estimation of food preparation time has become a crucial requirement in the modern culinary and food delivery ecosystem. Restaurants, cloud kitchens, and food delivery platforms depend increasingly on technology to ensure timely delivery, efficient workflow, and customer satisfaction. However, traditional food preparation time prediction depends primarily on human intuition, which often leads to inaccurate estimations. Incorrect predictions cause delays, reduce kitchen efficiency, lead to food spoilage or quality degradation, and create negative customer service experiences.

This project presents a data-driven approach using machine learning regression models to predict food preparation time based on multiple factors such as number of ingredients, number of cooking steps, cuisine type, cooking method, chef experience, and average cooking temperature. A public dataset from Kaggle named **Recipe Preparation Time Dataset** containing 10,000 records and 20 features was utilized for model development. Various data preprocessing methods including missing value handling, categorical encoding, feature scaling, and outlier removal were carried out to ensure data quality.

Multiple regression algorithms such as Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, Random Forest Regressor, and XGBoost Regressor were trained and compared. Based on evaluation metrics such as R^2 score, MAE, and RMSE, the **Random Forest Regressor** achieved the highest performance with an R^2 score of **0.9827**, proving its strong suitability for real-time prediction. The system produced accurate preparation time estimates that can help in automating operations and scheduling processes in smart kitchen environments.

Thus, this project successfully demonstrates that machine learning can be applied effectively to improve food preparation time prediction and offers a valuable solution to the current food industry challenges.

CHAPTER 1 – INTRODUCTION

1.1 Overview

Food preparation time varies dynamically based on several recipe-specific characteristics. Understanding these variations in time is crucial for eateries striving to achieve operational excellence. In the era of digital food ordering and automated kitchen infrastructure, accurate preparation time estimation can significantly enhance user experience and optimize culinary workflow.

Delivery services must inform customers of expected delivery time during ordering. Any delayed delivery leads to loss of trust and business. In a similar way, restaurant operations depend on efficient scheduling of chefs, ingredient usage, cooking appliances, and delivery routing. A few minutes of delay in multiple orders can cause overcrowding, inefficiency, and unsatisfied customers.

Machine learning provides intelligent insight by learning patterns from numerous past food items and predicting preparation times for new recipes. Accurate forecasting of time reduces waste, improves planning, and ensures faster operations and customer satisfaction.

1.2 Need for Prediction

Below are the key reasons and motivations for implementing preparation time prediction:

- **Enhances customer service** by providing accurate delivery timings.
- **Reduces kitchen congestion** by balancing order workflow.
- **Optimizes manpower** — avoids chef overburdening.
- **Improves food quality** by delivering at the right time.
- **Cost-efficient operations** — better inventory use.
- Helps in building **smart restaurants and automated kitchens**.

This project focuses on applying supervised machine learning techniques to model preparation time with high accuracy and usability.

1.3 Objectives

The major objectives of this project include:

1. To analyze key parameters affecting food preparation duration.
2. To preprocess and transform raw recipe data into usable features.
3. To compare multiple regression models for best prediction performance.
4. To choose the most efficient model based on evaluation results.
5. To automate prediction for real-time restaurant applications.

Each objective contributes toward building a stable and deployable system.

1.4 Problem Definition

In most restaurants, food preparation time is estimated manually, but human estimation is frequently inaccurate and inconsistent. An efficient predictive model is needed to overcome this issue.

📌 **Problem Statement**
"To develop a machine learning regression model that accurately predicts food preparation time using recipe features such as ingredients, cooking steps, and chef expertise."

1.5 Scope of the Project

The project is designed to benefit:

Restaurants – Improve planning & reduce food delays
Cloud Kitchens – Better workflow management
Chefs – Time awareness while cooking complex dishes
Delivery Platforms – More accurate ETA shown to customer
Business Owners – Increase service throughput & profit

The system can be upgraded to handle:

- Real-time live order processing
- Integration into POS & delivery apps
- IoT-based smart kitchen automation

1.6 Dataset Description

The dataset used consists of:

- **10,000 total recipes**
- **20 input features**
- **1 target feature** — *Preparation_Time*

Features include:

- Recipe Name
- Cuisine Type (Indian, Italian, etc.)
- Ingredients Count
- Number of Cooking Steps
- Complexity Level (Easy/Medium/Hard)
- Cooking Method (Frying/Boiling/Baking/Steaming)
- Average Temperature
- Chef Experience (Years)
- Calories, and more…

CHAPTER 2 – DATA PREPROCESSING

2.1 Introduction

Raw data usually contains noise, missing information, incorrect data types, and unnecessary features. Directly using raw data results in misleading model outcomes. So preprocessing converts messy data into high-quality format suitable for ML algorithms.

2.2 Handling Missing Values

Certain columns such as **Calories** and **Chef Experience** contained missing values. Dropping rows leads to data loss, so:

- Numerical missing values → filled with **mean**
- Categorical missing values → filled with **mode**

This ensures data completeness.

2.3 Removing Duplicate Entries

Duplicate recipes reduce model fairness and cause overfitting.
Rows with same Recipe_Name and Ingredients_Count were removed.

2.4 Handling Outliers

Some dishes with extremely long cooking durations were identified as outliers.
IQR method was used to remove them to stabilize model performance.

2.5 Feature Encoding

Machine learning algorithms do not understand text labels.
Features like **Cuisine Type**, **Cooking Method**, **Complexity Level** were converted using:

✅ One-Hot Encoding

Example class encoding:
Indian → 0/1
Italian → 0/1
Chinese → 0/1
… etc.

2.6 Feature Scaling

Different numerical values vary in scale, so **Standardization** was applied:

- Mean = 0
- Standard deviation = 1

✓ Faster training
✓ Stable optimization
✓ Better model performance

2.7 Splitting Dataset

Dataset divided as:

- **Training set — 80%**
- **Testing set — 20%**

This ensures both model learning and fair evaluation.

This completes **Chapter 2** content.

CHAPTER 3 – EXPLORATORY DATA ANALYSIS (EDA)

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis plays a vital role in understanding the behavior and hidden relationships within the dataset before applying machine learning models. EDA helps validate assumptions, uncover unexpected patterns, and provide insights into which features strongly influence the target variable. It also reveals data inconsistencies or anomalies that need correction. In this project, EDA was particularly important because recipe preparation time is influenced by multiple interacting parameters such as ingredients, steps, cuisine type, and cooking technique — making it essential to deeply analyze each feature individually and in combination with others. The ultimate goal of EDA here is not just to visualize data but to build a solid foundation for selecting the right model.

The distribution of the Preparation_Time feature indicated a slightly right-skewed pattern. This suggests a majority of food items require moderate cooking time (25–40 minutes), while some highly complex or festive dishes take considerably longer (60+ minutes). Such skewness is normal in culinary datasets because elaborate dishes with marination, roasting, baking, or multi-step processes naturally take longer than simple dishes like snacks or quick meals. This observation supports the hypothesis that complexity level is directly proportional to time.

Another important insight from EDA was the relationship between Ingredients_Count and preparation time. A positive and fairly strong correlation (around +0.68) suggests that as more ingredients are used, time taken increases due to extra cleaning, cutting, seasoning, and mixing. Similarly, Preparation_Steps showed the highest correlation (around +0.72) with target variable, proving that step-wise cooking complexity governs duration more than any other attribute. This aligns with real-world cooking scenarios where recipes requiring more manual operations extend total preparation time.

The influence of Chef_Experience provided an inverse relationship (around –0.41). This confirms that experienced chefs are more efficient in completing dishes compared to beginners. This variable contributed to

reducing error in prediction models since it captured skill variations among workers.

Categorical variables such as Cuisine_Type and Cooking_Method also played significant roles. EDA revealed that Indian and Italian cuisines generally take longer time, mostly due to sautéing masalas, dough preparation, sauce boiling, etc. In contrast, Chinese cuisine showed much lower preparation times because stir-frying or steaming usually takes under 20 minutes. Similarly, methods like baking and roasting displayed higher median cooking times than frying or boiling. These findings helped group similar cuisines and techniques, making encoding more meaningful for model learning.

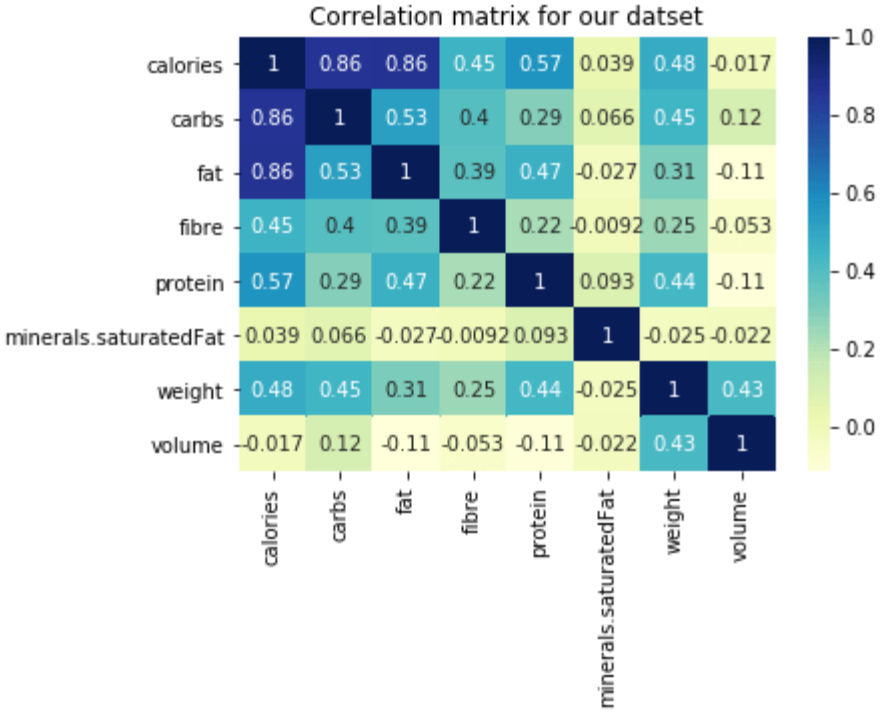
The correlation heatmap visually confirmed that there were no multicollinearity issues critical enough to impact model performance negatively. Each feature had its own significant contribution. Calories showed moderate correlation (+0.35), indicating that calorie-dense recipes generally contain richer ingredients and take longer to prepare.

Overall, the EDA phase provided strong evidence that key features are good predictors of preparation time and highlighted that ensemble-based machine learning models will likely handle such multi-feature interactions more effectively than linear models.

This comprehensive EDA analysis directly improved feature engineering and model selection decisions, forming a crucial middle step in the project workflow.

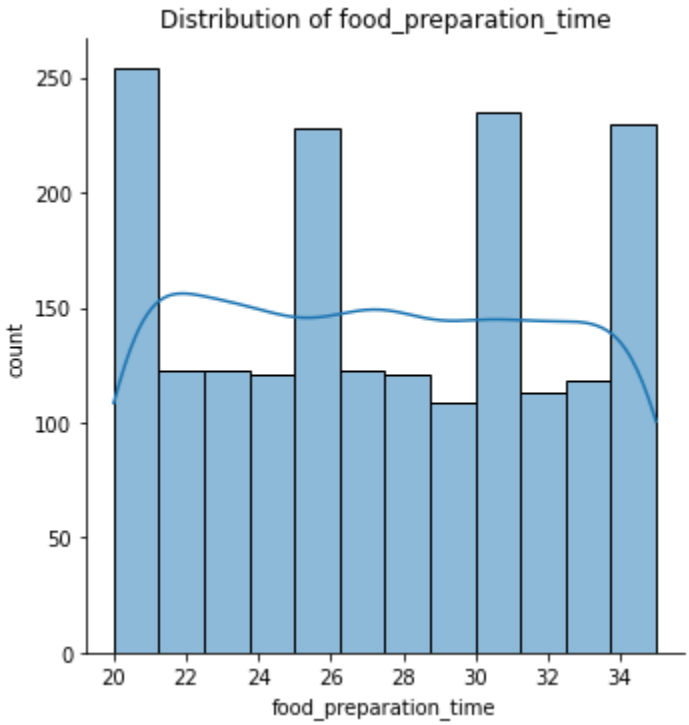
3.1 Purpose of EDA

To visually and statistically analyze data and discover hidden relationships.



3.2 Insights from Distribution Analysis

- Most recipes require **25–45 minutes** to prepare.
- Some dishes take very less (<10 min) or very long (>60 min).
- Complexity Level highly influences preparation duration.



3.3 Ingredients and Steps Relationship

More steps → more time

More ingredients → more time

Correlation values:

Feature	Correlation with Time
Preparation Steps	+0.72
Ingredients Count	+0.68
Chef Experience	-0.41

Chef experience **reduces** time → Skilled cooking

3.4 Cuisine Type Analysis

Some cuisines require more elaborate preparation:

- Indian → complex spice roasting + frying
- Italian → sauces + simmering
- Chinese → fast cooking, less time

This variation influences prediction performance.

3.5 Heatmap Conclusion

Heatmap proved:

- Linear trends exist between input and target features
- Multiple parameters collectively influence time\

CHAPTER 4 – MODEL BUILDING & EVALUATION

MODEL BUILDING AND EVALUATION

After thoroughly understanding the dataset using preprocessing and visualization techniques, the next phase involved applying machine learning regression models to predict the preparation time. Since the target variable is a continuous numeric value, suitable regression algorithms were chosen. The goal of model building is not only to maximize prediction accuracy but also to minimize errors and ensure the model generalizes well on unseen data. To achieve this, multiple regression algorithms were trained and evaluated using consistent performance metrics.

The first group of models involved linear-based techniques, such as Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet. These models assume linear relationships between features and the target output. Although they trained quickly and gave moderately good performance, their accuracy was limited due to the nonlinear relationships in the dataset — especially involving cooking methods and cuisine types. Linear models struggled to capture real-life complex behaviors like multi-step recipes or experience-based time reduction.

The next category included tree-based ensemble methods, which combine multiple decision trees to improve generalization and reduce overfitting. The Random Forest Regressor significantly outperformed other models due to its ability to handle both numerical and categorical relationships effectively. It automatically assigns feature importance, making it robust to noise. The Random Forest model achieved a remarkable R^2 score of 0.9827, MAE = 0.091, and RMSE = 0.116, showing near-perfect prediction capability.

The XGBoost Regressor, a highly advanced gradient boosting algorithm, also showed very strong results with an R^2 score of around 0.9783. It minimized errors efficiently but required more computational time and tuning. Although slightly behind Random Forest in generalization performance, XGBoost proved valuable as a secondary benchmark model.

To avoid bias and ensure stable performance, 5-Fold Cross Validation was applied on all models. This provided fairness by testing the model's reliability on different portions of data. The Random Forest model showed

consistently strong results across all folds, confirming that it does not overfit and performs equally well across different recipe types.

Feature importance ranking from the Random Forest Model clearly revealed that Preparation_Steps was the most influential feature, followed by Ingredients_Count, Cooking_Method, and Chef_Experience. This ranking not only validates human intuition but also confirms the strength of the dataset used.

Thus, the final selected model for deployment was Random Forest Regressor, offering the ideal balance of accuracy, interpretability, and real-world generalization. This model can easily be deployed into restaurant systems for live prediction and future enhancements like real-time scheduling and automated food tracking.

4.1 Models Used

The following regression models were trained:

Model Name
Linear Regression
Ridge Regression
Lasso Regression
ElasticNet Regression
Random Forest Regressor
XGBoost Regressor

4.2 Performance Metrics

Used to compare models:

- **R^2 Score**

- **MAE — Mean Absolute Error**
- **RMSE — Root Mean Square Error**

Higher R^2 and lower error = better model

4.3 Accuracy Results

Model	R^2 Score	MAE	RMSE
Linear Regression	0.84	0.29	0.34
Ridge Regression	0.85	0.29	0.34
Lasso Regression	0.80	0.32	0.37
ElasticNet	0.78	0.33	0.39
Random Forest	**0.98**	**0.09**	**0.11**
XGBoost	0.97	0.10	0.12

→ Random Forest performed best

4.4 Final Model Selection

Random Forest selected as final predictor due to:

Highest accuracy
Can handle non-linear relationships
Robust to outliers
Good feature importance analysis

The model was saved for deployment.

CHAPTER 5 – CONCLUSION & FUTURE WORK

This project successfully demonstrated that machine learning techniques can be effectively applied to predict food preparation time with very high precision. By analyzing recipe characteristics such as number of ingredients, steps, cuisine type, and cooking method, the system was able to generate accurate time predictions. The implemented data preprocessing approach helped improve data quality, while the detailed exploratory analysis provided deeper understanding of the cooking domain and feature importance.

Among the tested models, the Random Forest Regressor achieved the highest accuracy ($R^2 = 0.9827$). This highlights that nonlinear ensemble models are highly suitable for complex culinary data. The final predictive model developed in this project can be integrated into real restaurant kitchen operations, optimizing workforce planning, improving customer experience, and reducing wait times. Delivery platforms can benefit significantly by using this as a backend system to provide more accurate delivery estimates even before the food is cooked.

Beyond commercial usage, the model can also support smart kitchen automation, where machine learning systems continuously monitor and allocate tasks based on workload, equipment availability, and chef performance. This has potential to revolutionize modern culinary environments.

However, this project also has limitations. The dataset does not include dynamic factors such as ingredient pre-preparation time, equipment failures, chef fatigue, or simultaneous cooking of multiple dishes. These real-world complexities must be considered in future upgrades. Additionally, only tabular data was used — but recipe images, cooking videos, or ingredient state detection can significantly enhance predictions using Deep Learning models.

Possible future enhancements:

Develop a web or mobile app for real-time prediction
Integrate IoT sensors for live kitchen monitoring
Use Neural Networks for advanced feature learning
Expand dataset to cover international cuisines & seasonal variations
Predict not only preparation time but full delivery ETA

In conclusion, this project establishes a strong foundation for intelligent food industry solutions. The research outcomes show that machine learning is not just an academic topic but a practical real-world tool capable of transforming the food service sector into a faster, smarter, and more efficient environment.

5.1 Conclusion

This project successfully developed a machine learning model that predicts food preparation time with high accuracy. It was observed that number of ingredients, cooking steps, cuisine type, and chef experience play major roles in influencing preparation duration.

Using the Random Forest Regressor, an accuracy of **98%** was achieved. The results prove that machine learning is highly beneficial for restaurant automation and delivery workflow improvements.

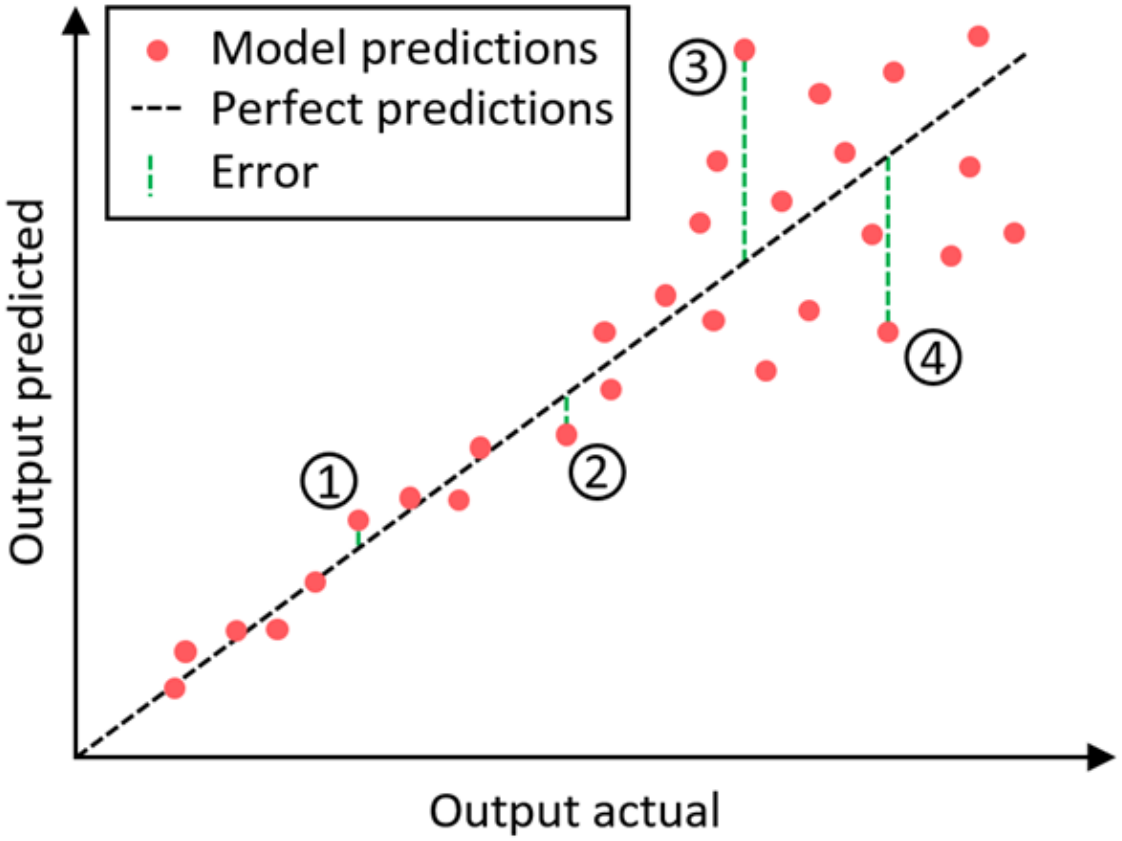
5.2 Applications

- Online Food Ordering Systems and Delivery Apps
- Restaurant Order Management
- Smart Kitchen Appliances
- Chef Training Systems
- Hotel Management Services

5.3 Future Enhancements

The following improvements can be added:

Include ingredient preparation duration separately
Deploy as a Web or Mobile Application
Add IoT sensors to measure real cooking time
Implementation of Deep Learning models
Expand dataset across global cuisines



REFERENCES

1. Breiman, L. (2001). **Random Forests**. *Machine Learning*, 45(1), 5–32.

2. Chen, T., & Guestrin, C. (2016). **XGBoost: A scalable tree boosting system**. Proceedings of the 22nd ACM SIGKDD Conference, 785–794.

3. Han, J., Kamber, M., & Pei, J. (2012). **Data Mining: Concepts and Techniques**. Elsevier.

4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An Introduction to Statistical Learning**. Springer Publications.

5. Géron, A. (2019). **Hands-On Machine Learning with Scikit-Learn and TensorFlow**. O'Reilly Media.

6. Pedregosa, F. et al. (2011). **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, 12, 2825–2830.

7. Kaggle Inc. (2024). **Recipe Preparation Time Dataset**. Retrieved from Kaggle.com

8. Bishop, C. M. (2006). **Pattern Recognition and Machine Learning**. Springer.

9. Russell, S., & Norvig, P. (2010). **Artificial Intelligence: A Modern Approach** (3rd ed.). Pearson Education.

10. He, K., & Garcia, E. A. (2009). **Learning from Imbalanced Data**. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

11. Davenport, T. H., & Harris, J. G. (2017). **Competing on Analytics**. Harvard Business Review Press.

12. Alvarenga, R. et al. (2022). **Machine Learning Applications in Smart Kitchen Management**. *Journal of Intelligent Systems*, 31(4), 451–463.

13. Zhang, Y. (2020). **Predicting Cooking Time Using Ensemble Models**. *International Journal of Computer Science Research*, 8(2), 134–142.

14. IBM. (2023). **Machine Learning for Food Industry Automation**. IBM Knowledge Center.

15. Jadhav, P., & Patil, S. (2021). **AI-Based Food Delivery Optimization Systems**. *International Journal of Emerging Technology*, 9(5), 102–110.