International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I5P106

Volume 73 Issue 5, 48-57, May 2025

Original Article

Utilizing Machine Learning and Feature Selection Techniques to Classify Osteoarthritis Risk Based on Consumption and Lifestyle Characteristics for the Population of Northern Thailand

Ploykwan Jedeejit[1], Wongpanya S. Nuankaew[2], Pratya Nuankaew[3]

[1]*Department of Marketing, Faculty of Business Administration, Bangkokthonburi University, Bangkok, Thailand.*
[2]*Department of Computer Science, School of Information and Communication Technology, University of Phayao, Phayao, Thailand.*
[3]*Department of Digital Business, School of Information and Communication Technology, University of Phayao, Phayao, Thailand.*

[3]*Corresponding Author : pratya.nu@up.ac.th*

Received: 31 January 2025 Revised: 22 April 2025 Accepted: 25 April 2025 Published: 31 May 2025

Abstract *- This research has three primary objectives: to study the environment and context of people's risk of osteoarthritis in the northern region of Thailand, to use machine learning and feature selection techniques to classify osteoarthritis risk based on consumption and lifestyle characteristics for the people in the north area of Thailand and to evaluate the model of classify osteoarthritis risk based on consumption and lifestyle characteristics for the people in the northern region of Thailand. The research data is a randomly selected sample of 351 representatives from eight villages in Ko Tan Subdistrict, Khanu Woralaksaburi District, Kamphaeng Phet Province, Thailand. The research instruments consisted of three questionnaires and nine machine-learning techniques. The research results are analyzed in two parts: descriptive statistical analysis and the efficiency of the machine learning model classification. The research results found that most of the sample were engaged in agriculture, aged between 61 and 70, and had primary education. The most efficient model is the random forest technique, which is 80 percent accurate, and seven significant factors affect osteoarthritis risk prediction.*

Keywords *- Applied Informatics, Medical Data Mining, Medical Informatics, Medical Innovations, Screening Osteoarthritis.*

1. Introduction

Osteoarthritis (OA) is a common degenerative joint disease that primarily affects older adults, causing chronic pain, mobility limitations, and a decreased quality of life. In Thailand, particularly among northern populations, the prevalence of OA is significantly influenced by local lifestyles and cultural practices. Activities such as squatting, carrying heavy loads, and climbing stairs are deeply embedded in daily routines and have been identified as potential contributors to knee joint degeneration [1]. Furthermore, dietary habits-such as low calcium intake, excessive consumption of fatty foods, and limited nutritional diversity-have also been associated with an increased OA risk [2]. Recently, Machine Learning (ML) has emerged as a powerful tool in healthcare analytics. It can analyse complex datasets and uncover hidden patterns that traditional statistical methods may overlook [3]. When combined with feature selection techniques, ML models can enhance predictive accuracy by identifying the most relevant variables from a large data pool, thereby improving model efficiency and interpretability [4]. Despite the growing body of research on OA prediction, few studies have focused on region-specific populations and their unique risk factors. Particularly in northern Thailand, there is still a gap in integrating localized dietary and lifestyle information into predictive models. Addressing this gap is essential for developing preventive strategies and early detection tools tailored to the specific needs and behaviours of these communities. Moreover, Thailand has a continuously increasing proportion of older adults, specifically those aged 60 years and over, while the proportion of Thai births has continually decreased [5-8].

This results in a shortage of working-age people to replace the aging population in the future. Another impact is that Thailand's social and economic structure has become unbalanced. However, many older adults in Thailand, where most of the population is poor, must continue working even after reaching 60. This situation leads to a significant increase

in health issues among the elderly. In 2023, Thailand had 13.64 million older adults (aged 60 years and over), accounting for 19.50 percent of the total population, thus qualifying Thailand as an aged society. An aged society is a community where 10 percent of the population is 60 years and over, or 7 percent is 65 years and over [9].

At the same time, the occupational situation of the elderly in Thailand was found to be 5.11 million people who continued to work after the age of 60, which is 37.46 percent of the total elderly population in Thailand. We classified the elderly workers into two groups: males, comprising 2.77 million individuals, accounting for 20.31 percent of older people, and females, comprising 2.34 million individuals, representing 17.16 percent of older people. Furthermore, when classifying the elderly workers by industry sectors, it was found that the majority of them worked in the agricultural sector (59.30%), the service and goods sector (30.50%), and the manufacturing sector (10.20%), which means that the majority of them still had to use physical labour in their occupations (National Statistical Office, Ministry of Digital Economy and Society, Thailand) [16].

Moreover, the National Statistical Office, Ministry of Digital Economy and Society, Thailand, reports that many problems affect elderly workers. The survey revealed that issues related to older people's work could be classified into three main categories: work problems, work safety problems, and work environment problems, as detailed below: 1) Work-related problems indicated that 823,000 older adults faced work-related issues, with 52.7 percent experiencing insufficient compensation for living, followed by employment discontinuity at 16.4 percent and overwork problems at 16.1 percent, among others. 2) Problems related to occupational safety showed that 492,000 elderly workers encountered this issue. At 75.6 percent, it involved exposure to various chemicals, followed by dangerous and unprotected machinery and tools at 14.8 percent, and effects on body organs at 3.1 percent, among others. Issues stemming from the working environment indicated that 654,000 elderly workers were affected by work postures, air pollution, dust, and waste in the workplace. The conclusions drawn have many health implications for older people across various dimensions [16].

In addition to its many effects, osteoarthritis is the most common issue among older individuals. In Thailand, 50 percent of people over 65 years old have osteoarthritis, and the tendency is increasing continuously, rendering it one of the critical public health challenges. If individuals with osteoarthritis do not receive treatment or do not take appropriate action, the symptoms of the disease will continue to progress. These conditions may lead to pain, deformity of the knee joints, and abnormal walking, making daily activities inconvenient and resulting in physical and mental suffering. Public health data indicates that osteoarthritis has increased, particularly in knee joints. This trend is attributed to the currently growing global aging trend. The World Health Organization predicts that the number of osteoarthritis patients in 2050 among individuals over 60 years old may increase by up to 15 percent, equating to approximately 130 million people suffering from osteoarthritis, especially affecting the knee joints.

This situation aligns with Thailand, where the incidence of osteoarthritis patients has also been continuously increasing. According to bone and joint disease statistics, in 2010, there were over 6 million patients with osteoarthritis, making it the most common cause of disability. Additionally, osteoarthritis is among the ten leading causes of disability in the elderly in Thailand.

Due to the importance of various factors, researchers have conducted this research with the following research objectives.

1.1. Research Objectives

This research has three primary objectives:

1. To study the environment and context of people's risk of osteoarthritis in the northern region of Thailand.
2. To use machine learning and feature selection techniques to classify osteoarthritis risk based on consumption and lifestyle characteristics for the people in the northern region of Thailand.
3. To evaluate the model of classifying osteoarthritis risk based on consumption and lifestyle characteristics for the people in the northern region of Thailand.

2. Literature Review

Osteoarthritis (OA) is a prevalent and progressive musculoskeletal disorder that affects millions globally, particularly the aging population. It commonly impacts weight-bearing joints such as the knees, leading to chronic pain, stiffness, and reduced mobility. In Thailand, studies have highlighted a growing burden of OA, especially among the elderly in rural and agricultural regions, such as the northern provinces, where lifestyle and occupational behaviors differ significantly from urban counterparts [10].

Numerous studies have explored the relationship between lifestyle factors and OA risk. The researchers [11] reported that habitual squatting, kneeling, and physically intensive agricultural work are prominent contributors to knee OA among northern Thai elders. Likewise, Kosulwat (2002) [12] found that dietary habits—such as low intake of calcium-rich foods, high saturated fat consumption, and poor nutritional diversity—can aggravate the risk and severity of OA. These findings emphasize the importance of region-specific risk assessment models.

In recent years, machine learning (ML) has gained traction as a promising approach in medical diagnosis and risk prediction. ML algorithms such as decision trees, support

vector machines (SVM), and random forests have shown excellent performance in classifying OA severity and predicting OA onset based on clinical, behavioural, and imaging data.

Crucially, the use of feature selection techniques enhances ML performance by reducing data dimensionality, eliminating irrelevant or redundant features, and improving model interpretability. Guyon and Elisseeff [4] introduced several feature selection methods such as filter, wrapper, and embedded techniques, each offering distinct advantages depending on the dataset size and model complexity. When applied to health data, feature selection helps identify key predictors of OA, enabling the development of efficient and interpretable diagnostic tools.

However, the majority of existing ML-based OA research has been conducted in Western contexts using clinical or biomechanical data, with minimal focus on lifestyle and dietary factors relevant to Southeast Asia. There is a clear research gap in integrating culturally and regionally specific lifestyle characteristics—such as food consumption patterns, postural habits, and environmental exposure—into predictive models for OA risk in northern Thailand.

This study seeks to bridge that gap by developing a machine learning classification model that incorporates both dietary and lifestyle variables tailored to the northern Thai population. The inclusion of feature selection techniques aims to optimize model accuracy and provide insights into the most influential factors contributing to OA risk within this demographic.

3. Materials and Methods

3.1. Population and Sampling

The researcher selected the population and sample using the statistical method based on Krejcie and Morgan's principles, selecting people from eight villages in Ko Tan Subdistrict, Khanu Woralaksaburi District, Kamphaeng Phet Province, Thailand, as detailed in Table 1.

Table 1. Population and sampling

Village Name	Population	Samples with Equal Proportion
Baan Rong Sub	802	71
Baan Ko Tan	594	52
Baan Rai Dong Yang	344	30
Baan Non-Tarot	646	57
Baan Don Khaen	329	29
Baan Khok Sa-at	403	36
Baan Don Khwang	353	31
Baan Nong Kham	514	45
Total	3,985	351

Table 1 illustrates the population and sample data in the research. The researchers used Krejcie and Morgan's statistical selection method, which used the equal proportion sampling method calculated from the population divided by the sample size (3,958/351 = 11.28). Once the proportion of the population to be sampled was determined, it was calculated for each village.

3.2. Data Collection

The researchers conducted data collection in the target communities. They collected data from individuals using the Thai language for communication and asked permission to collect data via questionnaires, as indicated in Tables 2 to 5. The activities and data collection process are presented in Figures 1 to 2.

Fig. 1 Data-gathering activities (scenario 1)

Fig. 2 Data-gathering activities (scenario 2)

3.3. Research Tools

The research instrument is a questionnaire consisting of five parts as follows:

Part 1 is the collection of demographic data.

Part 2 is the risk of osteoarthritis assessment with the Oxford Knee Score [13, 14], as detailed in Tables 2 to 3.

Part 3 is the survey of daily food consumption behaviour, as detailed in Table 4.

Part 4 is the survey of lifestyle behaviour, as detailed in Table 5.

Part 5 is the respondents' recommendations for this research.

Table 2. The Oxford Knee Score (OKS) english version

Questions and Options
During the past 4 weeks.....
1. How would you describe the pain you usually have from your knee? ❒ None (4) ❒ Very mild (3) ❒ Mild (2) ❒ Moderate (1) ❒ Severe (0)
2. Have you had any trouble washing and drying yourself (all over) because of your knee? ❒ No trouble at all (4) ❒ Very little trouble (3) ❒ Moderate trouble (2) ❒ Extreme difficulty (1) ❒ Impossible to do (0)
3. Have you had any trouble getting in and out of a car or using public transport because of your knee? (whichever you would tend to use) ❒ No trouble at all (4) ❒ Very little trouble (3) ❒ Moderate trouble (2) ❒ Extreme difficulty (1) ❒ Impossible to do (0)
4. How long have you been able to walk before the pain from your knee becomes severe? (with or without a stick) ❒ No pain/More than 30 minutes (4) ❒ 16 to 30 minutes (3) ❒ 5 to 15 minutes (2) ❒ Around the house only (1) ❒ Not at all - pain severe when walking (0)
5. After a meal (sitting at a table), how painful has it been for you to stand up from a chair because of your knee? ❒ Not at all painful (4) ❒ Slightly painful (3) ❒ Moderately painful (2) ❒ Very painful (1) ❒ Unbearable (0)
6. Have you been limping when walking because of your knee? ❒ Rarely/never (4) ❒ Sometimes, or just at first (3) ❒ Often, not just at first (2) ❒ Most of the time (1) ❒ All of the time (0)
7. Could you kneel down and get up again afterwards? ❒ Yes, Easily (4) ❒ With little difficulty (3) ❒ With moderate difficulty (2) ❒ With extreme difficulty (1) ❒ No, Impossible (0)
8. Have you been troubled by pain from your knee in bed at night? ❒ No nights (4) ❒ Only 1 or 2 nights (3) ❒ Some nights (2) ❒ Most nights (1) ❒ Every night (0)
9. How much pain from your knee has interfered with your usual work (including housework)? ❒ Not at all (4) ❒ A little bit (3) ❒ Moderately (2) ❒ Greatly (1) ❒ Totally (0)
10. Have you felt your knee might suddenly 'give way' or let you down? ❒ Rarely/never (4) ❒ Sometimes, or just at first (3) ❒ Often, not just at first (2) ❒ Most of the time(1) ❒ All of the time (0)
11. Could you do the household shopping on your own? ❒ Yes, Easily (4) ❒ With little difficulty (3) ❒ With moderate difficulty (2) ❒ With extreme difficulty (1) ❒ No, Impossible (0)
12. Could you walk down one flight of stairs? ❒ Yes, Easily (4) ❒ With little difficulty (3) ❒ With moderate difficulty (2) ❒ With extreme difficulty (1) ❒ No, Impossible (0)

Table 2 presents the Oxford Knee Score (OKS) osteoarthritis risk assessment items and questions. The questionnaire has five options and five scores, which were used to calculate the severity level, as shown in Table 3.

Table 3. The Oxford Knee Score (OKS) criteria

Score range	Score grading	Severity level
0 – 19	May indicate severe knee osteoarthritis.	Level 4
20 – 29	May indicate moderate to severe knee osteoarthritis.	Level 3
30 – 39	May indicate mild to moderate knee osteoarthritis.	Level 2
40 – 48	May indicate satisfactory joint function.	Level 1

Table 4. The food consumption behavior questionnaire

Stage	Questions for food consumption behavior
FC01	Consistency in consuming high-fiber foods such as eggplant, passion fruit, guava, etc.
FC02	Consistency in consuming low-fiber foods such as cucumber, lychee, longan, etc.
FC03	Consistency in fast food consumption.
FC04	Consistency in consuming fried foods.
FC05	Consistency in consuming 6 – 8 glasses of water per day.
FC06	Consistency in consuming soft drinks and syrups.
FC07	Consistency in consuming tea and coffee.
FC08	Consistency in consuming alcohol.
FC09	Consistency in consuming milk.
FC10	Consistency in consuming instant noodles and

	canned foods.
FC11	Consistency in consuming snacks.
FC12	Consistency in consuming fermented foods.
FC13	Consistency in consuming spicy foods.

Table 5. The lifestyle behavior questionnaire

Stage	Questions for lifestyle behavior
LS01	How often do you use the squat toilet?
LS02	How often do you sit on the floor for thirty minutes continuously?
LS03	How often do you sit and work in a chair for eight hours straight?
LS04	How often do you stand for more than three hours at work?
LS05	How often do you walk to work for three hours straight?
LS06	How often do you get up from your desk every hour?
LS07	How often do you go up and down more than three flights of stairs per day?
LS08	How often have you continuously held or carried objects for more than ten minutes?
LS09	How often do you exercise for thirty minutes or more continuously?
LS10	How often do you do vigorous exercises that involve jumping or jogging?
LS11	How often do you stretch before exercising?
LS12	How often do you do housework for at least three hours straight?
LS13	How often do you go outdoors with equipment such as a cane or a wheelchair?

Table 4 shows questions about food consumption behaviour consisting of 13 questions. Table 5 displays questions on lifestyle composed of 13 questions with five levels (scores) of information criteria: Level 1 contains a value of 0 points, meaning it has never been consumed. Level 2 includes a value of 1 point, which means it has been consumed 1 – 2 times per week. Level 3 contains a value of 2 points, meaning it has been consumed 3 – 4 times weekly. Level 4 comprises a value of 3 points, which means it has been consumed 5 – 6 times per week. Level 5 retains a value of 4 points, meaning it has been consumed regularly daily.

3.4. Model Construction

The model construction uses the CRISP-DM [15-17] data mining theory and is comprised of six stages.

3.4.1. Business Understanding

Business understanding involves understanding the problem and issue, which includes the research question. This research aims to study the impact of consumption and lifestyle behaviour on the risk of osteoarthritis by classifying various behaviours using descriptive statistical analysis techniques and machine learning classification techniques.

3.4.2. Data Understanding

Understanding data involves reaching the core and essence of the data. The data used in this research is sensitive and personal, and the research team has processed it through the research ethics process.

This process requires asking for permission from the informants to consent to use the data for research purposes. Before collecting data, the research supervisor and data collectors always questioned the informants.

3.4.3. Data Preparation

Data preparation is the longest step, and it includes data collection. This research is a primary data collection control, in which the researcher has to collect the data himself. In addition, the target source is rural people, who are mostly old and have vision problems.

Moreover, most of the informants have completed basic education, and some have not, which makes data collection difficult. After collecting data, researchers prepared it in a format suitable for developing a risk classification model for osteoarthritis.

3.4.4. Modeling

The development process of this model uses nine machine-learning techniques [17], including Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. After model development, each model is tested to determine model performance.

3.4.5. Evaluation

The evaluation step determines the model's performance. To construct and test the model, researchers split the data into two parts: the training dataset, which has 80 percent of the total data, and the test dataset, which has 20 percent of the total data. Four metrics-accuracy, precision, recall, and F1, which are known as confusion metrics-are required to evaluate the model's performance.

3.4.6. Deployment

Deployment is the process of applying the results and making the research tangible. Researchers have planned the application development to facilitate and create valuable products for the community.

3.5. Research Analysis and Interpretation

The analysis and interpretation of the research results are divided into two parts. The first part is interpreting the results for descriptive statistics analysis, in which the researcher uses max (maximum), min (minimum), mode, median, mean, and other essential techniques. The second part is the performance analysis of the model, which involves dividing the test data and using confusion matrix analysis techniques.

4. Results and Discussion

4.1. Research Results

The research findings are categorized into two sections. The first section presents the research findings through descriptive statistics in the analysis, as illustrated in Tables 6 to 12. The second section presents the findings from developing the model for classifying osteoarthritis risk, utilizing seven machine learning techniques and the model's performance, as detailed in Tables 13 and 14.

The collected samples were classified into 232 patients with osteoarthritis (66.10%) and 119 general persons (33.90%). Of these, 180 subjects (51.28%) were male, and 171 subjects (48.82%) were female. All details are shown in Table 6. Table 7 shows the sample data collected classified by age. Most of the sample was between 61 and 70 years old, with 76 cases or 21.65 percent, followed by those between 51 and 60 years old, with 59 cases or 16.81 percent. The smallest group was 10-20 years old, with three people, or 0.85 percent.

Table 6. Examples of data classified by gender

Gender	Osteoarthritis Patients	General Persons	Total
Male	109 (31.05%)	71 (20.23%)	180 (51.28%)
Female	123 (35.04%)	48 (13.68%)	171 (48.52%)
Total	232 (66.10%)	119 (33.90%)	351 (100.00%)

Table 7. Examples of data classified by age

Age	Osteoarthritis Patients	General Persons	Total
10 – 20 yrs.	3 (0.85%)	0 (0.00%)	3 (0.85%)
21 – 30 yrs.	1 (0.28%)	36 (10.26%)	37 (10.54%)
31 – 40 yrs.	14 (3.99%)	31 (8.83%)	45 (12.82%)
41 – 50 yrs.	33 (9.40%)	21 (5.98%)	54 (15.38%)
51 – 60 yrs.	44 (12.54%)	15 (4.27%)	59 (16.81%)
61 – 70 yrs.	68 (19.37%)	8 (2.28%)	76 (21.65%)
71 – 80 yrs.	37 (10.54%)	5 (1.42%)	42 (11.97%)
81 – 90 yrs.	27 (7.69%)	3 (0.85%)	30 (8.55%)
More than 91 years.	5 (1.42%)	0 (0.00%)	5 (1.42%)
Total	232 (66.10%)	119 (33.90%)	351 (100.00%)

Table 8. Examples of data classified by occupation

Occupation	Osteoarthritis Patients	General Persons	Total
State Enterprise Employees	0 (0.00%)	3 (0.85%)	3 (0.85%)
Private Employees	3 (0.85%)	20 (5.70%)	23 (6.55%)
Traders/Private Businesses	28 (7.98%)	9 (2.56%)	37 (10.54%)
Farmers	101 (28.77%)	26 (7.41%)	127 (36.18%)
General Employment	27 (7.69%)	27 (7.69%)	54 (15.38%)
Students	3 (0.85%)	18 (5.13%)	21 (5.98%)
Housewives	6 (1.71%)	1 (0.28%)	7 (1.99%)
Unemployed	64 (18.23%)	15 (4.27%)	79 (22.51%)
Total	232 (66.10%)	119 (33.90%)	351 (100.00%)

Table 8 shows the sample data collected classified by occupation. Most of the sample were farmers, with 127 cases, or 36.18 percent, followed by the unemployed, with 79 cases, or 22.51 percent. The smallest group stated enterprise employees, with 3 people, or 0.85 percent.

Table 9 shows the sample data collected classified by education. Most of the sample were primary school respondents, with 190 cases, or 54.13 percent, followed by lower secondary school respondents, with 50 cases, or 14.25 percent. The smallest group was comprised of bachelor's degree respondents, with 30 cases, or 8.55 percent.

Table 9. Examples of data classified by education

Education	Osteoarthritis Patients	General Persons	Total
Primary School	130 (37.04%)	60 (17.09%)	190 (54.13%)
Lower Secondary School Level	39 (11.11%)	11 (3.13%)	50 (14.25%)
Upper Secondary School Level	22 (6.27%)	25 (7.12%)	47 (13.39%)
Diploma	23 (6.55%)	11 (3.13%)	34 (9.69%)
Bachelor's Degree	18 (5.13%)	12 (3.42%)	30 (8.55%)
Postgraduate Degree	0 (0.00%)	0 (0.00%)	0 (0.00%)
Total	232 (66.10%)	119 (33.90%)	351 (100.00%)

Table 10. Examples of data classified by OKS assessment results

OKR Assessment Result	Absolute Count	Faction
Level 4 (Severe)	183	0.521
Level 3 (Moderate)	125	0.356
Level 2 (Mild)	30	0.085
Level 1 (Normal)	13	0.037
Total	351	1.00

Table 10 shows the examples of data classified by OKS assessment results. Most of the sample were level 4, with 183 cases, or 52.14 percent, followed by level 3, with 125 cases, or 35.61 percent. The smallest group was level 1, with 13 cases, or 3.70 percent.

Table 11. Analysis results of the food consumption behavior questionnaire

Stage	Consistency of Consumption					Weight	S.D.
	(4)	(3)	(2)	(1)	(0)		
FC01	30	57	160	94	10	2.01	0.94
FC02	36	40	151	119	5	1.95	0.96
FC03	13	30	30	179	99	1.09	1.02
FC04	2	61	99	183	6	1.63	0.81
FC05	348	2	1	0	0	3.99	0.13
FC06	30	17	58	191	55	1.36	1.07
FC07	135	13	57	74	72	2.19	1.60
FC08	44	37	47	87	136	1.33	1.40
FC09	80	57	107	106	1	2.31	1.14
FC10	35	56	39	159	62	1.55	1.23
FC11	30	20	65	134	102	1.26	1.18
FC12	5	33	41	220	52	1.20	0.85
FC13	46	102	89	71	43	2.11	1.22

Table 11 presents the results of the analysis of the consumption behaviour of the collected samples. It was found that the behaviour of the samples was diverse. However, the factor with the highest weight was "FC05: consistency in consuming 6 – 8 glasses of water per day", with a weight of 3.99 and S.D. = 0.13. It reflects that the sample group is aware of their health care.

Table 12. Analysis results of the lifestyle behavior questionnaire

Stage	Consistency of Consumption					Weight	S.D.
	(4)	(3)	(2)	(1)	(0)		
LS01	4	13	42	99	193	0.68	0.90
LS02	0	6	12	141	192	0.52	0.65
LS03	10	10	31	78	222	0.60	0.96
LS04	6	13	29	61	242	0.52	0.92
LS05	52	49	35	75	140	1.42	1.49
LS06	157	110	18	53	13	2.98	1.20
LS07	18	4	13	147	169	0.73	0.98
LS08	22	4	18	121	186	0.73	1.06
LS09	10	2	79	112	148	0.90	0.96
LS10	7	21	11	152	160	0.75	0.92
LS11	8	130	93	98	22	2.01	1.00
LS12	93	223	28	1	6	3.13	0.70
LS13	10	10	5	18	308	0.28	0.86

Table 12 presents the results of the analysis of the lifestyle behaviour of the collected samples. The behaviour of the samples was also diverse. The lifestyle behaviour with the highest weight is factor LS12, "How often do you do housework for at least three hours straight?" with a weight of 3.13 and S.D. = 0.70. This means that most lifestyle factors influence the development of osteoarthritis.

Table 13. Model analysis with machine learning

Model	Accuracy	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Naive Bayes	73.22%	7.94%	104	419943	346.34	146.34
Generalized Linear Model	67.54%	4.68%	92	465591	1365.85	207.32
Logistic Regression	69.13%	8.16%	94	443338	1075.61	420.73
Fast Large Margin	76.12%	4.49%	110	485086	1085.37	1182.93
Deep Learning	82.07%	1.65%	124	460577	1309.76	310.98
Decision Tree	69.24%	3.44%	96	466969	609.76	243.9
Random Forest	84.57%	4.08%	130	517248	704.88	1506.1
Gradient Boosted Trees	79.46%	3.78%	118	551672	2117.07	1207.32
Support Vector Machine	79.35%	3.26%	118	522254	873.17	701.22

Table 13 shows the results of model development and performance analysis. The random forest technique had the highest accuracy, with an accuracy of 84.57% and an S.D. of 4.08%. The model's performance analysed by the confusion matrix technique is shown in Table 14, and the significant factors are shown in Table 15.

Table 14. Model performance

	True Level 1	True Level 2	True Level 3	True Level 4	Class Precision
Pred. Level 1	27	3	0	1	87.10%
Pred. Level 2	2	17	6	3	60.71%
Pred. Level 3	0	0	25	0	100.00%
Pred. Level 4	0	3	0	30	90.91%
Class Recall	93.10%	73.91%	80.65%	88.24%	

Table 15. Random forest weights

Stage	Attribute	Weight
LS12	Do housework	0.4147
LS11	Stretch before exercising	0.1237
LS08	Holding an object for more than 10 minutes	0.0924
LS09	Exercise for more than 30 minutes	0.0831
FC07	Consuming tea and coffee	0.0802
FC13	Consuming spicy foods	0.0733
FC04	Consuming fried foods	0.0617

Table 15 shows the model weights reflecting the significant attributes of the feature selection techniques.

4.2. Research Discussion

The discussion of the research results is divided into three parts that are in line with the research objectives.

4.2.1. Environment and Context of People's Risk of Osteoarthritis in the Northern Region of Thailand

Tables 6 to 12 present the results of the summary and analysis of the context and environment of the collected samples. It can be summarized as follows: The majority of the sample is male, with 180 instances (51.28%), and aged 61-70, with 76 cases (21.65%). Moreover, most occupations are farmers, with 127 instances (36.18%), and the education level of the sample is at the lowest level, primary school, with 190 respondents (54.13%). In addition, the evaluation of the four severity levels of the OKS criteria found that most of the sample had the highest level and risk of osteoarthritis (Level 4: Severe), with 183 cases (52.10%). The next rank is the third level (Level 3: Moderate), with 125 numbers (35.60%). While the asymptomatic and mildly ill group was in the smaller group, they included Level 1: Normal, with 13 instances, and Level 2: Mild, with 30 cases. The results show that the sample group in this research is at a high risk of severe osteoarthritis. Therefore, promoting care and finding ways to prevent it is necessary and appropriate.

4.2.2. Classification Osteoarthritis Model

This research applied machine learning and data mining techniques to complete a risk prediction model for osteoarthritis based on people's consumption and lifestyle behaviours in Northern Thailand. The model performance analysis results are presented in Table 13. It was found that the random forest technique has the highest accuracy, with 84.57 percent, a reasonable level of standard deviation, with 4.08 percent, and the best scoring time, with 1506.1 per 1,000 rows. Therefore, it is sensible to conduct a thorough performance test of this model and determine its significant factors.

4.2.3. Performance of the Osteoarthritis Classification Model and its Significant Factors

Table 13 shows the model analysis results, which found that the random forest technique had the highest accuracy value when tested with the confusion matrix technique, as reported in Table 14. The researchers found that all performance evaluation indicators were high, including Level 1's recall of 93.10%, Level 2's recall of 73.91%, Level 3's recall of 80.65%, Level 4's recall of 88.24%, Level 1's precision of 87.10%, Level 2's precision of 60.71%, Level 3's precision of 100.00%, and Level 4's precision of 90.91%. In addition, the researchers found seven significant factors affecting the random forest model, as shown in Table 15.

Table 15 presents the significant factors for the random forest model, which are as follows. The most significant factor was item LS12: "How often do you do housework for at least three hours straight?" which weighed 0.4147. The second most significant factor was item LS11: "How often do you stretch before exercising?", which weighed 0.1237. The third most significant factor was item LS08: "How often have you continuously held or carried objects for more than ten minutes?" which weighed 0.0924. The fourth most significant factor was item LS09: "How often do you exercise for thirty minutes or more continuously?" which weighed 0.0831. The fifth most considerable factor was item FC07: "Consistency in consuming tea and coffee.", which weighed 0.0802. The sixth most significant factor was item FC13: "Consistency in consuming spicy foods.", which weighed 0.0733. The last significant factor was item FC04: "Consistency in consuming fried foods.", which weighed 0.0617. The results and discussions reflect the success of the research in achieving the objectives set and the study plan.

5. Conclusion

The setting of the Thai population transitioning into an aging society, characterized by over 10 percent of individuals aged sixty and older residing in the community, underscores the necessity of establishing healthcare provisions for this substantial demographic segment. This research seeks to develop technology to reduce healthcare expenses for Thailand's aging population. The primary objectives are threefold: to examine the environmental and contextual factors influencing the risk of osteoarthritis in northern Thailand, to employ machine learning and feature selection methodologies to categorize osteoarthritis risk based on the consumption and lifestyle attributes of the northern Thai population, and to assess the classification model of osteoarthritis risk derived from these consumption and lifestyle characteristics in the northern region of Thailand.

The population and sample for this research comprised meticulously chosen target groups from eight villages in the Ko Tan Subdistrict, Khanu Woralaksaburi District, Kamphaeng Phet Province, Thailand, as outlined in Table 1. The sample size was determined using Krejcie and Morgan's formula, resulting in a total of 351 samples, also detailed in Table 1. The research instruments are categorized into two primary sections: the data collection tool, utilizing questionnaires for data acquisition, and the second category, which encompasses tools for data mining model development

and machine learning analysis, employing nine techniques for model development and the confusion matrix for performance evaluation.

The analysis results were satisfactory and in line with the objectives set. The essential features of the sample group's context included that most participants were farmers, with 127 instances (36.18%), and that they possessed the lowest level of education, with 190 replies (54.13%). Furthermore, most respondents were unaware of their susceptibility to developing osteoarthritis, as the OKS questionnaire evaluation indicated. Table 10 shows that most respondents were classified at the highest risk level (Level 4: Severe), comprising 183 instances (52.10%). Therefore, there should be a tracking system to safeguard and care for this target population.

The researchers employed nine machine learning techniques to create the osteoarthritis risk prediction model, with the analysis results presented in Table 13. The random forest technique demonstrated the best accuracy at 84.57%, so it was chosen to evaluate its performance using the confusion matrix method, as illustrated in Table 14. Seven key elements affected the development of this model, as indicated in Table 15. The researchers earnestly hope this study will prove beneficial and contribute to future implications in Thailand.

5.1. Limitation

A notable limitation of this research is that it is an undergraduate student project that encourages students to learn and explore research experiences. Therefore, some steps have slight glitches. However, all researchers tried their best to achieve all the objectives of this research.

Funding Statement

Three organizations-the Thailand Science Research and Innovation Fund, the Bangkokthonburi University, and the University of Phayao-supported this research project.

Acknowledgments

This research project also received support from many advisors, academics, researchers, students, and staff. The authors thank everyone for their support and cooperation in completing this research.

References

[1] Umaporn Hanrungcharotorn et al., "Factors Influencing Physical Activity among Women with Osteoarthritis of the Knee," *Pacific Rim International Journal of Nursing Research*, vol. 21, no. 1, pp. 5-17, 2017. [Google Scholar] [Publisher Link]

[2] Ampicha Nawai et al., "Factors Associated with Nutrition Risk Among Community-Dwelling Older Adults in Thailand," *Geriatric Nursing*, vol. 42, no. 5, pp. 1048-1055, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ziad Obermeyer, and Ezekiel J. Emanuel, "Predicting the Future-Big Data, Machine Learning, and Clinical Medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216-1219, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[4] Isabelle Guyon, and André Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003. [Google Scholar] [Publisher Link]

[5] Sutthichai Jitapunkul, and Suvinee Wivatvanit, "National Policies and Programs for the Aging Population in Thailand," *Ageing International*, vol. 33, no. 1, pp. 62-74, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[6] Patcharawalai Wongboonsin, Yupin Aungsuroch, and Naomi Hatsukano, *The Ageing Society and Human Resources to Care for Older People in Thailand*, Human Resources for the Health and Long-Term Care of Older Persons in Asia, pp. 104-135, 2020. [Google Scholar] [Publisher Link]

[7] Pramote Prasartkul, Suchada Thaweesit, and Sutthida Chuanwan, "Prospects and Contexts of Demographic Transitions in Thailand," *Journal of Population and Social Studies*, vol. 27, no. 1, pp. 1-22, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shotaro Kumagai, "Thailand's Efforts to Cope with a Rapidly Aging Population," *RIM Pacific Business and Industries*, vol. 19, no. 71, pp. 2-28, 2019. [Google Scholar] [Publisher Link]

[9] Srawooth Paitoonpong, and Natcha Yongphiphatwong, "Promotion of Active Aging and Quality of Life in Old Age and Preparation for a Complete Aged Society in Thailand," *TDRI Quarterly Review (September 2023)*, vol. 38, no. 3, pp. 1-35, 2024. [Google Scholar] [Publisher Link]

[10] Weena Phungviwatnikul, and Jutharath Voraprateep, "Analysis of Aging Society in Thailand between A.D. 2016-2022," *Journal of Arts and Thai Studies*, vol. 46, no. 1, pp. 1-13, 2024. [Google Scholar] [Publisher Link]

[11] Prin Vathesatogkit et al., "Associations of Lifestyle Factors, Disease History and Awareness with Health-Related Quality of Life in a Thai Population," *PLOS ONE*, vol. 7, no. 11, pp. 1-9, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[12] Keerati Charoencholvanich, and Boonchana Pongcharoen, "Oxford Knee Score and SF-36: Translation & Reliability for Use with total Knee Arthroscopy Patients in Thailand," *Journal-Medical Association of Thailand*, vol. 88, no. 9, pp. 1194-1202, 2005. [Google Scholar] [Publisher Link]

[13] J.-Y. Jenny, and Y. Diesinger, "The Oxford Knee Score: Compared Performance Before and After Knee Replacement," *Orthopaedics & Traumatology: Surgery & Research*, vol. 98, no. 4, pp. 409-412, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[14] Sarah L. Whitehouse et al., "The Oxford Knee Score; Problems and Pitfalls," *The Knee*, vol. 12, no. 4, pp. 287-291, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mihaela Cazacu, and Emilia Titan, "Adapting CRISP-DM for Social Sciences," *BRAIN: Broad Research in Artificial Intelligence and Neuroscience*, vol. 11, no. 2Sup1, pp. 99-106, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Fernando Martínez-Plumed et al., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048-3061, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526-534, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Daniel, Tjeng Wawan Cenggoro, and Bens Pardamean, "A Systematic Literature Review of Machine Learning Application in COVID-19 Medical Image Classification," *Procedia Computer Science*, vol. 216, pp. 749-756, 2023. [CrossRef] [Google Scholar] [Publisher Link]